8-28-2019

The Use of a Targeted Omic Approach to Dissect the Nested Symbiosis of the Eastern subterranean termite, Reticulitermes Flavipes

Michael Stephens
University of Connecticut - Storrs, michael.stephens@uconn.edu

The Use of a Targeted Omic Approach to Dissect the Nested Symbiosis of the Eastern subterranean termite, *Reticulitermes flavipes*

Michael Eric Stephens Jr. PhD

University of Connecticut 2019

The Eastern subterranean termite, *Reticulitermes flavipes* is a member of lower wood-feeding termites and feeds on lignocellulose (wood) in nature. The ability to survive on this hard-to-digest, nutrient-poor diet relies on a division of labor between the host termite and its hindgut microbiota. Key players of this symbiotic digestive strategy include a community of various protist species and their associated, prokaryotic, endo- and ectosymbionts. These protists aid the hydrolysis of the various polysaccharides found in wood and it is thought that their bacterial symbionts contribute to a number of processes that are regarded as essential in the termite's hindgut. The inability to currently culture these organisms has hindered our ability to understand these complex interactions and their contributions to the digestion of wood.

By focusing on these protist-associated communities of bacteria and leveraging the sensitivity and coverage of high-throughput sequencing, this complex hindgut community can be dissected and studied to reveal ecological, metabolic, and evolutionary processes that have previously evaded us. This body of work, described in this thesis, aides in our understanding of these communities by first describing the community structures and member transmission trends and then investigating their gene content and expression of both endo- and ectosymbionts. These data have led to precise hypotheses about these protist-associated communities, some of which I was able to test and support. For example, endosymbiotic *Endomicrobium* species have acquired

many genes related to their carbon metabolism by horizontal gene transfer, perhaps through a competence pathway which was both conserved and expressed.

Due to the complexity of this hindgut community these targeted analyses, and future extensions of similar types, are crucial to our pursuit of understanding of the underlying processes that occur in the termite's hindgut and how they relate to the efficient digestion of wood.

The Use of a Targeted Omic Approach to Dissect the Nested Symbiosis of the Eastern subterranean termite, *Reticulitermes flavipes*

Michael Eric Stephens Jr.

B.S., University of North Florida, 2012

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2019

APPROVAL PAGE

Doctor of Philosophy Dissertation

The Use of a Targeted Omic Approach to Dissect the Nested Symbiosis of the Eastern subterranean termite, *Reticulitermes flavipes*

Presented by

Michael Eric Stephens Jr., B.S.

Major Advisor

_____

Daniel J. Gage

Associate Advisor

_____

Joerg Graf

Associate Advisor

_____

Johann Peter Gogarten

Associate Advisor

_____

Robertson Thane Papke

Associate Advisor

_____

Spencer Nyholm

University of Connecticut
2019

Acknowledgements

I would first like to thank Dr. Daniel Gage for his mentorship and guidance over these last six years. He taught me how to think critically about my research, as well as dedicated a lot of time towards understanding and troubleshooting the omic approaches that we used in these studies. I could not have done this without him. I would also like to thank all the members of my thesis committee, Dr. Joerg Graf, Dr. Spencer Nyholm, Dr. Peter Gogarten, Dr. Thane Papke, and Dr. Jonathan Klassen who set aside time to meet with me over these last six years and helped refine my research by both challenging me and directing me towards better experimental designs. I cannot thank them enough.

I would also like to thank the other members of the lab, Charles Bridges, Jamie Micciulla, and Gabrielle Corso, who have been not only great colleges but also amazing friends. I would also like to thank all the past lab members who encouraged me during the beginning of my graduate studies. I would also like to thank all the undergraduate researchers who I had the pleasure of working with during my time here at UConn. I also would like to thank all of my peers who began this journey with me back in 2013, especially Andrea Suria who has been a great friend over this last six years. I couldn't have done this without her continued support and insights.

I would like to thank my family who have always encouraged and supported my childhood dream of becoming a "mad scientist". They taught me to value education and to pursue my passion, even if they didn't quite understand what I was passionate about. I also would like to thank my fiancé André Beringhs, who has been by my side over the last several years and has offered constant support and encouragement.

Table of Contents

List of Tables

Chapter Three

Chapter Four

List of Figures

Chapter Two

Chapter Three

Chapter Four

Chapter One

Introduction

1.1 General Termite Biology and Research Interests

Termites are eusocial insects that belong to the phylogenetic clade Blattodea along with their cockroach relatives (Evangelista et al., 2019). Divergence of termites from their roach ancestors occurred approximately 100 million years ago during the Cretaceous Period (Evangelista et al., 2019). Since then, termites have had managed to become extremely successful in terms of their distribution (occupying every continent except Antarctica) and large population sizes which in some cases results in their collective biomass rivaling that of larger co-inhabitants, such as mammals (Brune, 2014). Their success is surprising, given that all termite species feed on nutritionally poor lignocellulose, which is a component of the cell wall of woody plants (Brune, 2014). However, their ability to efficiently degrade lignocellulose combined with their eusociality are thought to be the main contributors to their ecological success (Brune, 2014).

Although all termites feed on lignocellulose in various forms, only a subset of termite species feed on wood. Their feeding activities and colony structures have profound effects on the ecosystems in which they are found. For example, the activities of termites have been found to increase crop yields by as much as 36% in dry agricultural settings by aiding in the infiltration of water through soil and also increasing the nitrogen content in those soils (Evans et al., 2011). Furthermore the feeding activities of termites collectively accounts for 1-3 % of global methane emissions, even though most of the methane produced by termites is consumed within their colonies or mounds prior to escaping into the atmosphere (Nauer et al., 2018). However, some termite species which feed on wood are regarded as pest insects as they pose a threat to wooden

infrastructures with termite related damages collectively costing on the order of billions of US dollars per year (Su and Scheffrahn, 1998). For example, the feeding activity of an invasive Asian termite *Copotermes gestori*, is currently causing damage to both wooden infrastructure as well as the urban tree canopy in southeastern Florida (Chouvenc and Foley, 2018).

Termites have long fascinated biologists not only because of their complex interactions with microbial symbionts, which are housed in their hindgut, but also due to the demand to develop novel pest control strategies (Brune, 2014; Scharf, 2014). In recent years, the use of Omic technologies have been used to generate targeted termite control strategies which include the development of termiticides, RNA interference (RNAi), and engineered organisms to combat the threat of pest termite species (Scharf, 2014). However, the ability to efficiently degrade lignocellulose in the form of wood and other varieties is mostly studied in the context of gaining a better understanding of the underlying processes involved, such that they can be applied to the development of novel biotechnologies for sustainable energy (Scharf, 2014, 2015). Since approximately 50-65% of the composition of lignocellulose is carbohydrate, it represents most abundant renewable energy source on Earth (Scharf, 2014). However, our current understanding of the mechanisms involved in the digestion of lignocellulose, which is considered resistant to enzymatic degradation, is limited but growing (Brune, 2014; Scharf, 2014). The hurdles associated with the digestion of such a food source and how termites overcome these challenges are discussed below.

1.2 Digestion of Wood and Symbiotic Cooperation in Termites

Wood-feeding termites are charged with the task of overcoming two main hurdles associated with feeding exclusively on lignocellulose. First, since lignocellulose is a complex of various

polysaccharides such as cellulose, hemicelluloses, pectin, and lignin it is regarded as being resistant enzymatically hydrolysis and is therefore hard to digest. Second, lignocellulose (in the form of wood) is regarded as being nutrient poor since it is void of essential nitrogenous compounds and other sources of nitrogen (Brune, 2014). Thus, termites must both efficiently degrade the polysaccharides found in wood, as well as, supplement their diet with essential nitrogenous compounds which are otherwise absent or limiting.

Phylogenetically basal lineages of termites (referred to as lower termites) overcome these challenges by dividing the labor of polysaccharide hydrolysis between their own mechanical and chemical processes with the hydrolytic capacity of microbial symbionts which live in their hindguts (Brune, 2014). In addition, these microbial symbionts carry out essential nutritional functions within the hindgut to ensure that their termite hosts acquire both necessary carbon and nitrogen sources (Brune, 2014; Ohkuma et al., 1996; Pester and Brune, 2006). The digestion of wood by termites requires symbiotic cooperation between the insect host and various microbial partners.

Polysaccharide degradation begins with the mechanical processing of wood by the termite's mandibles which yield small pieces that are further processed by the termite's gizzard (Brune, 2014). These small bits are mixed with termite salivary enzymes in the midgut, which begin to hydrolyze some of the polysaccharides like cellulose, freeing sugar monomers which are absorbed by the midgut epithelium (Brune, 2014). However, it is thought that the majority of the wood remains undigested in the midgut and is passaged into the hindgut lumen where a dense community of various flagellated protist species phagocytize the wood particles and complete polysaccharide hydrolysis (Brune, 2014). Then, short chain fatty acids such as acetate, which are

end products of microbial fermentation, are absorbed by the termite hosts and used as their primary carbon source (Brune, 2014).

Nitrogen enters the termite hindgut in the form of atmospheric N_2 gas which diffuses across the termite's tissues and is converted into ammonia by the nitrogenase activity of nitrogen fixing bacteria (Brune, 2014; Ohkuma et al., 1996). The ammonia is then incorporated into biomass initially by bacteria and then by protists which may feed on those bacteria (Brune, 2014). A portion of the hindgut microbial biomass is voided by termites in hindgut fluid which they then feed to other members of their colony to recover essential nitrogenous compounds (Brune, 2014). Those nitrogenous compounds are then metabolized by the termite hosts and waste, in the form of uric acid, is secreted back into the hindgut where is it recycled by uricolytic bacteria (Brune, 2014).

The cooperation between lower termites, protists, and prokaryotic symbionts in their hindguts are responsible the efficient digestion of lignocellulose. However, details of this symbiotic digestive strategy remain unknown. Currently, it is not known (i) what the contributions of individuals or even groups of certain organisms are nor (ii) the exact molecular mechanism for some of these processes.

1.3 Protist and Bacterial Symbionts of Termite Hosts

The protist community of wood feeding lower termites belong to two eukaryotic taxa which include the Parabasalia (phylum or class) and Oxymonadida (order). These symbionts are housed in the termite's hindgut where they occupy the majority of the hindgut volume and live either in the luminal space or are attached to the hindgut wall. Several lines of evidence support that these symbionts contribute to polysaccharide hydrolysis in the hindgut and are essential to their termite

hosts. First, removing protist symbionts from their termite hosts results in the inability of the termites to continue to live while feeding on wood (Brune, 2014). Second, previously cultivated protists isolated from termites were shown to degrade cellulose in vitro (Trager, 1934; Yamin and Trager, 1979). Finally, wood particles are commonly seen in the cytoplasm of many of these termite protist species supporting that wood can be used as a food source.

Enzymes of protist origin have experimentally been shown to efficiently degraded some of the polysaccharides founds in wood (Sethi et al., 2013). These include enzymes belonging to glycosyl hydrolase family 7 (GHF7, also known as exoglucanases), which work synergistically to degrade cellulose to glucose with GHF9 (endoglucanases) and GHF1 (β-glucosidases) enzymes (Brune, 2014; Sethi et al., 2013). GHF9 enzymes randomly cleave β-1,4-glycosidic bonds of amorphous regions of cellulose which releases chains of cellulose fibers from wood (Brune, 2014; Sethi et al., 2013). Then GHF9 enzymes cleave those cellulose chains from the ends to produce cellobiose which are then hydrolyzed by the GHF1 enzymes into glucose (Brune, 2014; Sethi et al., 2013).

The protist symbionts of termites also represent an ecological niche which is commonly colonized by bacteria. Most protists that live in lower termites are colonized by both endo- and ectosymbionts (Ohkuma, 2008). It is thought that the bacterial symbionts of protist participate in some of the essential metabolic and nutritional functions within the termite hindgut such as the assimilation of essential nitrogenous compounds and the recycling of fermentation end products into acetate (Ohkuma, 2008). However, since these organisms are not yet cultivated and are hard to isolate from their termite hosts, the symbioses of protists and bacteria are not well understood.

Previous studies have investigated these symbioses by either obtaining genomes for some of these bacterial symbionts or identifying them through fluorescence in situ hybridization (FISH).

Here, I will briefly describe some these associations which pertain to protists symbionts of *Reticulitermes* spp. termites. Attached to the plasma membrane of protist hosts by one cell pole are bacterial ectosymbionts which belong to with the genera *Treponema* and '*Candidatus* Symbiothrix'(Hongoh et al., 2007; Iida et al., 2000; Radek and Tischendorf, 1999). These two types of ectosymbionts co-colonize protists hosts, in some cases, and live in intermixed population with one another. The populations of ectosymbiotic *Treponema* are composed of at least two types, which cluster together into two distinct phylogenetic groups known as the Termite *Treponema* Clusters I & II (Iida et al., 2000). Although currently there are no genomic data on these *Treponema* it is hypothesized that that they are acetogens which recycle protist fermentation end products, as well as, fix nitrogen (Ohkuma, 2008). These hypotheses originate from the observation that some cultivated *Treponema* isolates, from other termites, carry out those functions in culture (Graber and Breznak, 2004; Lilburn et al., 2001). The draft genome for '*Ca.* Symbiothrix' suggests that this particular ectosymbiont may participate in polysaccharide hydrolysis since its genome is enriched in glycosyl hydrolases (Yuki et al., 2015).

The endosymbionts which colonized the cytoplasm of termite protists include a few different bacterial taxa, however genome analysis from a few of these members suggests that despite differences in their phylogeny, their genomes have retained similar gene content. These include members of the bacterial genus *Endomicrobium*, which in *Reticulitermes* spp. colonize the cytoplasm of both Parabasilia and Oxymonadida protist hosts (Hongoh et al., 2008a; Stingl et al., 2005). In other termite species, the endosymbionts of protists also include member of '*Candidatus* Azobacteroides' (Hongoh et al., 2008b), '*Candidatus* Ancillula' (Strassert et al., 2016), and *Treponema* (Ohkuma et al., 2015). Overall, these endosymbionts retain complete pathways to assimilate essential amino acids and vitamins and it is hypothesized that the protist

hosts may recover these compounds by digesting some of their endosymbionts over time. Other associations between protists and bacteria include intranuclear symbionts which are members of the Verrucomicrobia, embedded ectosymbionts which are members of the *Desulfovibrio*, and Margulisbacteria which live on the surfaces of some ectosymbiotic *Treponema*.

Bacteria also occupy other niches within the hindguts of termites which include the hindgut wall, as well as, the viscous fluid of the lumen. Unlike most environments on Earth, the hindguts of lower termite are often dominated by *Treponema* which in addition to being protist-associated, also represent the majority of free-living bacteria (Benjamino and Graf, 2016; Boucias et al., 2013). In *Reticulitermes flavipes*, *Treponema* represent approximately 40% of the core bacterial community which is shared across different termite colonies (Benjamino and Graf, 2016). Other members of the hindgut bacterial community include Firmicutes, and various members of the Proteobacteria, Synergistes, and Bacteroidetes. Interestingly, most of the bacterial phyla which are present in the core bacterial community of *R. flavipes* contain members that are protist-associated, however the proportion of free-living to protist-associated bacteria within those phyla are not known. Since the hindgut protists occupy the majority of the hindgut volume, they likely influence the composition of the bacterial community.

1.4 An Overview of the Use of Omics to Study Symbioses

Omic approaches are invaluable to the study of symbioses as most members of a given microbial community cannot yet be cultivated and culture-based experiments usually require that organisms be studied out of context of their associations (Nayfach and Pollard, 2016a). Omics offers, in a culture-independent manner, the ability to extract biological materials (DNA, RNA, protein, or metabolites) from a given environment such that the composition of

those materials can be readily assayed in the context of the interactions within that environment (Douglas, 2018; Nayfach and Pollard, 2016a). The data generated by Omic approaches can serve as types of molecular observations, from which precise hypotheses can be developed and in many cases, these hypotheses can be tested using other experimental approaches (Douglas, 2018; Scharf, 2015).

One such example of a widely used Omic approach to investigate symbioses are high-throughput amplicon surveys, which usually target conserved marker genes for taxonomic purposes (Caporaso et al., 2012). In these studies, the taxonomic diversity and distribution of microbial symbionts across tissue types or host organisms, can be readily obtained. Examples of the use of this approach span many well-studied symbioses, which include hosts such as the Hawaiian Bobtail Squid (*Euprymna scolopes*) (Kerwin and Nyholm, 2018), fungus growing ants (*Trachymyrmex septentrionalis*) (Ishak et al., 2011), and termites (*Reticulitermes flavipes*) (Benjamino and Graf, 2016). In these studies, a variable region the bacterial 16S rRNA gene is amplified and then sequenced using a high-throughput sequencing platform such as the Illumina MiSeq, which provides thousands of reads per sample of that single genetic locus (Caporaso et al., 2012).

Recent advances in high-throughput amplicon analysis have allowed for the investigation into the taxonomic diversity of microbes, within and across samples, at the resolution of single nucleotides. These techniques account for the sequencing error within reads to derive the true biological sequence of the genetic locus of interest (Callahan et al., 2017). These are referred to as amplicon sequence variants (ASVs), exact sequence variants (ESVs), and oligotypes (Callahan et al., 2016, 2017; Eren et al., 2013). These analytical approaches offer several advantages over alternative methods such as clustering reads into operational taxonomic units

(OTUs), at or above an arbitrary dissimilarity threshold, to themselves (de novo) or reference sequences (reference-based). One advantage to these approaches is that they are not reliant of reference databases, which cannot truly account for the biological diversity of any given genetic locus in nature (Callahan et al., 2017). These approaches are also more reproducible across studies since they derive true biological sequences and not subject to artifacts due to clustering algorithms (Callahan et al., 2017). True biological sequences can be found and compared across different studies and environments, whereas clustering reads may give different OTUs depending on the method used (reference-based or de novo) (Callahan et al., 2017). Recently, these approaches have been shown to more accurately depict the composition of mock bacterial communities compared to clustering methods (Caruso et al., 2019).

Beyond taxonomic data, Omic approaches offer other avenues to investigate aspects of symbioses such as metabolic properties of individuals, as inferred by their gene content and gene expression. The use of shotgun metagenomics and metatranscriptomics offer a culture-independent genomic survey of members in a microbial community (Nayfach and Pollard, 2016). Using these methods, total DNA and RNA of a given environment can be sequenced without the need to separate or cultivate individuals from that environment (Nayfach and Pollard, 2016). Generally, these data are analyzed either through read-based approaches or assembly-based approaches in which either reads or assembled contigs are annotated by the function of the genes in which they encode, as well as, a putative taxonomic assignment for the organism from which they originated (Huson et al., 2016; Nayfach and Pollard, 2016; Quince et al., 2017).

Some recent advances in the analytical tools used to process metagenomic data have made it possible to recover near complete genomes from metagenomic assemblies (Laczny et al., 2015; Sangwan et al., 2016). These are often referred to as draft genomes, genomic bins, or

metagenome-assembled genomes (MAGs). The ability to assemble and recover MAGs has recently contributed to our understanding of the phylogenetic diversity of many yet to be cultivate organisms (Parks et al., 2017). For example, nearly 8,000 MAGs were recovered in a single study using over 1,500 metagenomes (Parks et al., 2017).

Metagenomics approaches have previously been applied to the study of termites, focusing on the microbial communities of the termite hindguts (Scharf, 2015). In general, these studies have identified genes related to essential hindgut functions and in some cases, identified the groups responsible for those functions (Liu et al., 2019; Tokuda et al., 2018). For example, a recent metagenomic study of a wood-feeding higher termite (*Nasutitermes takasagoensis*) revealed that *Treponema* symbionts carry out hemicellulase activity in that termite's hindgut (Tokuda et al., 2018). However, in lower termites which have a more diverse microbial hindgut community and more microenvironments (such as the cytoplasm and surfaces of protist cells), the assignment of function based on metagenomic data is more complicated. The hurdles with dissecting such a complex community full of inmate interactions between symbionts require a more direct approach, such that the context of the associations between protist hosts and their bacterial symbionts are not lost. The approaches to overcome such hurdles are described below and represent the scope of this dissertation.

1.5 Scope of Dissertation

This dissertation is a collection of studies in which I obtain both taxonomic and functional data from the bacterial symbionts of individual protist hosts, which were isolated from the hindgut of the Eastern subterranean termite (*R. flavipes*). The use of isolated single protist cells has allowed me to perform a molecular dissection of this complex community by

leveraging both high-throughput amplicon as well as, metagenomic and metatranscriptomic sequencing.

I will begin by describing the bacterial communities of single celled hosts, focusing on three predominant protist hosts in *R. flavipes*, which are members of the genera *Trichonympha*, *Pyrsonympha*, and *Dinenympha*. Overall these protists associated with a distinct bacterial community which were relatively conserved across different termite colonies. However, the *Dinenympha* species, which associated with the most diverse bacterial communities, shared several *Treponema* symbionts which led to the hypothesis that at least some of these ectosymbiont can be horizontally acquired. This hypothesis was experimentally tested and supported using a novel fluorescence assay to track members of the hindgut community over time.

Next, I will describe my work related to using single (host) cell shotgun metagenomics and metatranscriptomics to recover near complete genomes of protist-associated bacteria. In these studies, I assembled the near complete genomes of three novel *Endomicrobium* species, which are the endosymbiont of some protists. Genome analyses revealed several interesting differences between the endosymbionts of *Trichonympha*, *Pyrsonympha*, and *Dinenympha* hosts including differences in carbon and nitrogen metabolisms as well as aerotolerance. These genomes also acquired some of the genes in those metabolic pathways by horizontal gene transfer (HGT) from putative donor taxa which are present in the hindguts of termites. Further analysis revealed that all *Endomicrobium* genomes have retained a putative competence pathway and that the endosymbionts have both conserved and expressed these genes while residing in their protist host cell. The expression of a key gene *comEC* by *Endomicrobium,* was verified by reverse transcriptase PCR (RT-PCR) of protist fraction

cDNA. Collectively these data support that these endosymbionts are naturally competent and may import and recombine exogenous DNA.

Finally, I will conclude my dissertation work by describing assembly and read-based approaches used to investigate the gene content and expression of ectosymbiotic *Treponema* populations. With these data I have focused on determining their binding proteins which may be involved in attachment to protist hosts, and gene content and expression data related to essential hindgut function such as reductive acetogenesis and nitrogen fixation. Collectively these data supported that the ectosymbionts of different protist hosts share similar expression profiles and that transcripts for metabolic pathways related to essential hindgut functions are present in each of the *Treponema* communities. Experimentally I will show that, in *R. flavipes*, protein mediates attachment of ectosymbionts to protist hosts and using the Omic data, I propose certain genes as candidates for those encoding the proteins involved in that binding mechanism. These gene candidates can be targets for further study such that their role in binding can be more precisely investigated.

The microbial communities of lower termites like *R. flavipes* are complex compared to the gut microbiome of other insects. They include members of all three domains of life, most of whom, cannot yet be cultivated. Their intimacies with one another further complicates their understanding by introducing nested levels of interactions (bacterial symbionts of protist hosts who are themselves, gut symbionts of termites). However, by targeting single protist cells for high-throughput sequencing, this complex community can be further dissected. The use of Omics approaches in this dissertation has shown that these molecular data can be used to derive specific hypotheses, pertaining to yet-to-be cultivated microbial symbionts, which can be experimentally tested. These are necessary approaches that can be

expanded and approved upon, in our pursuit to understanding the nested symbioses of lower termites.

Chapter Two

Amplicon Sequence Variant (ASV) Inference Reveals Community Structures and Transmission Trends of Protist-Associated Bacteria in the Termite *Reticulitermes flavipes*[+*]

[+]Under review, Stephens and Gage 2019
*Preprint of an earlier version of the manuscript is available on the bioRxiv Pre-Print server (https://doi.org/10.1101/299255) (Stephens and Gage, 2018).

Abstract

The hindgut protists of wood-feeding termites are usually colonized by prokaryotic endo- and ectosymbionts. Although these associations are thought to play critical roles in the efficient digestion of lignocellulose, they are not well understood. Many of the hurdles that have prevented a better understanding of these symbioses arise from variation between protist and termite host species and the inability to maintain prominent microbial members in culture. These issues have made it difficult to assess the fidelity, acquisition, and differences in colonization of bacterial symbionts by individual protist hosts. In this study, we use high throughput amplicon sequencing of V4 16S rRNA genes to determine the composition of bacterial communities associated with single protist cells of three predominate genera of hindgut protists (*Pyrsonympha*, *Dinenympha*, and *Trichonympha*) from the hindgut of the termite *Reticulitermes flavipes*. Using amplicon sequence variant (ASV) inference, the diversity and distribution of protist-associated bacteria was compared within and across six different protist species from the three genera. In general, each protist genus associated with a distinct community of bacterial symbionts which were conserved across different termite colonies. However, some ASVs were found to be shared between different protist species. For example, ASVs corresponding to ectosymbionts (Spirochaetes) were shared between different *Dinenympha* species and to a lesser extent with *Pyrsonympha* and *Trichonympha* hosts. This suggested that certain bacterial symbionts may be shared and perhaps acquired by their protist hosts by horizontal transmission. Using a fluorescence-based cell assay, we observed the horizontal transmission of ectosymbionts over time. Ectosymbiont transmission was shown to be dependent on time, involve active processes, and was non-random with respect to binding locations on some host cells. Although some symbionts were horizontally acquired there are likely to be as yet unidentified mechanisms in place to maintain their specificity to certain range of protist hosts. Our results suggest that

ASV inference of bacteria associated with single protist cells, combined with cell-based transmission assays, can reveal insights into the complexity of their bacterial symbionts' community structures and their host associations.

Introduction

The lower termite *R. flavipes* harbors symbionts from the three domains of life, all of which make significant contributions to the digestion of lignocellulose and provisioning of essential nutrients. These symbionts include uncultivated hindgut protists of two eukaryotic taxa, Oxymonadida (order) and Parabasalia (phylum or class) (Brune, 2014; Ohkuma, 2008). Most, and perhaps all, of these protists are colonized by both endo- and ectosymbionts from various bacterial taxa (Ohkuma, 2008). These protist-associated bacteria often exhibit complex community structures and occupy different ecological niches on and within their unicellular host (Ohkuma, 2008).

Previous studies have shown that Oxymonadida protists in *Reticulitermes speratus* are co-colonized with *Treponema* ectosymbionts from two distinct phylogenetic clusters (Termite *Treponema* clusters I and II) (Iida et al., 2000; Noda et al., 2003) as well as a member of the *Bacteroidales* (Hongoh et al., 2007). These three ectosymbionts lineages attach by one cell pole to the plasma membrane of their host (Radek and Tischendorf, 1999) and exhibit intermixed colonization (Hongoh et al., 2007). Other known ectosymbionts include a *Desulfovibrio* species which embeds in the membrane of its Parabasalia host, *Trichonympha* (Sato et al., 2009). Functional or genomic data regarding the nature of these symbioses are limited but growing. Genome analysis of a *Bacteriodales* ectosymbiont of a *Dinenympha* species found in *R. speratus*, '*Candidatus* Symbiothrix dinenymphae', suggests that it may directly degrade lignocellulose and/or aid its protist host in cellulose degradation (Yuki et al., 2015). In addition, a recent study

described ectosymbiotic 'Margulisbacteria' ('*Candidatus* Termititenax') which were attached to the surface of *Treponema* ectosymbionts of some Oxymonadida species in different termite hosts (Utami et al., 2019).

The endosymbionts which colonize the cytoplasm of hindgut protists of *Reticulitermes* termites are also composed of several bacterial taxa which vary between different protist and termite species. These include *Endomicrobium* (Stingl et al., 2005), and '*Candidatus* Ancillula' (Strassert et al., 2016). In addition, the nuclei of some hindgut protist species are colonized by *Verrucomicrobia* (Sato et al., 2014). Genome analysis of some of these endosymbionts suggests that there is convergent evolution for these symbionts to support their unicellular host by synthesizing nutrients absent in the termites' diet (Hongoh et al., 2008b, 2008a; Strassert et al., 2016). Regarding *Endomicrobium*, previous studies have investigated both their population structure and transmission in *Trichonympha* spp. protist hosts. Across various *Trichonympha* spp. these endosymbionts share rRNA gene phylogenies that are congruent with their host (Ikeda-Ohtsubo and Brune, 2009), are composed of a single phylotype, and are not shared across different *Trichonympha* spp. (Zheng et al., 2015). These data supported the hypothesis that these endosymbionts are vertically acquired.

The associations between termite protists and their symbiotic bacteria are only beginning to be understood. For example, in *Reticulitermes* spp. different protist species associate with *Treponema* from the same phylogenetic clusters, but the composition and fidelity of those associations are not resolved beyond those broad phylogenetic groups. Furthermore, assessing the diversity of bacteria which associate with termite protists has been challenging since these protists are not yet cultivated, and are hard to isolate from their termite hosts. Previous studies have overcome these challenges by either using samples which consisted of pooled protist cells

(Noda et al., 2003) or single-cell samples in which whole genome amplification (WGA) was performed (Zheng et al., 2015). These methods have yielded novel information, but with recent advances in high-throughput amplicon sequencing and analysis we wondered if we could assay single-cells without the need the need to pool individuals or perform WGA.

Here, we used a method in which uncultivated single protist cells, isolated from the hindgut of *R. flavipes*, served as a template for high-throughput amplicon sequencing of the hypervariable V4 region of bacterial symbiont 16S rRNA genes as well as co-amplifying the 18S rRNA gene of some of the protist hosts. Using this method, the bacterial symbiont composition of single protist cells can be investigated at single nucleotide resolution and high coverage. In addition, the ability to co-amplify the 18S rRNA gene of individual host cells gives further insight into these complex interactions by determining the phylogenetic relationship of individual protist cells. We selected six different protist species, four from the genus *Dinenympha* (Oxymonadida), *Pyrsonympha vertens* (Oxymonadida), and *Trichonympha agilis* (Parabasalia). Using these protist species, we investigated whether closely related protists, living in the same termite host, associated with similar bacterial symbionts compared to more distantly related protists. Using single isolated protist cells from the hindgut of *R. flavipes* ASVs were generated for protist-associated bacteria and then compared across different protist species from several termite colonies. The use of ASV inference provided single nucleotide resolution into the diversity of bacterial taxa associated with single protist cells (Callahan et al., 2016, 2017). Our results support previous observations in other *Reticulitermes* spp. termites that some hindgut protists share ectosymbiotic *Treponema*.

The observation that different protists species share some ectosymbiotic ASVs suggests that these particular symbionts may be cosmopolitan and horizontally acquired by their protist

hosts. To test this hypothesis, we developed an *in vitro* fluorescence assay which allowed us to detect the horizontal acquisition of ectosymbionts by different protists species. We show that the horizontal transmission of these ectosymbionts required active biological processes and the symbionts exhibited preferential spatial binding to their host cells in some cases. Using this high-resolution molecular approach combined with our transmission assay, we show most symbiont ASVs are exclusive to either Parabasalia and Oxymonadida protists in *R. flavipes* and that their community structures vary with respect to their ecological niche and mode of transmission.

Materials and methods

Termite collection, maintenance, and identification

R. flavipes termites were collected from three separate colonies using cardboard traps placed under logs for 2 to 4 weeks in Mansfield and Granby Connecticut, USA (Colony A: Longitude -72.262216, Latitude 41.806543; Colony B: Longitude -72.789053, Latitude 41.999316; Colony C: Longitude -72.216138, Latitude 41.759949). Termites were removed from traps and maintained in plastic containers with moistened sterile sand and spruce wood. Species identity of the termites was verified to be *R. flavipes* by solider caste morphology (Ye et al., 2004), the presence of *Dinenympha gracilis* in the hindguts of worker termites (Lewis and Forschler, 2004, 2006), and sequencing of the cytochrome oxidase II gene (Supplementary Figure 1) with primers A-tLEU: 5'- ATGGCAGATTAGTGCAATGG-3' (forward) and B-tLys: 5'-GTTTAAGAGACCAGTACTTG-3' (reverse)(Maekawa et al., 1999). For all experiments, only individuals of the worker caste from three termite colonies were chosen.

Amplification and sequencing of protist and bacterial SSU rRNA genes

Samples consisting of single protist cells were prepared from termites in an anaerobic chamber with atmospheric content of CO_2 5.5%, H_2 5.5%, and N_2 89%. Hindguts were dissected

and then ruptured in Trager's solution U (TU) (Trager, 1934). Hindgut contents were then washed by centrifugation at 3,000 rpm for 90 seconds, and then resuspended in TU for a total of three washes. Samples were then diluted and spotted on glass slides treated with RNase AWAY® Reagent (Life Technologies) and UV light. Individual protist cells were isolated using a micromanipulator (Eppendorf CellTram® Vario) equipped with a hand-drawn glass capillary. Protists cells were washed three times in droplets of TU via micromanipulation, placed in 10µl molecular grade TE Buffer, and then frozen at -20°C. In total, we collected and analyzed 57 protist cells from 13 worker termites which were sampled across three different termite colonies.

Frozen protist cells served as templates for PCR reactions in which the 18S rRNA gene of the protist host as well as the V4 hypervariable region of the 16S rRNA gene of bacteria were co-amplified and sequenced. PCR reactions consisted of Phusion® High-fidelity polymerase (1 unit), HF buffer, dNTPs (200 µM), dimethyl sulfoxide (DMSO) (3%), 0.3 µM of each 18S primer (Euk19f, 5'-AYYTGGTTGATYCTGCCA-3' and Euk1772r; 5'-CBGCAGGTTCACCTAC-3') (Ohkuma et al., 1998), 0.2 µM each of V4 16S primers (515f; 5'-GTGCCAGCMGCCGCGGTAA-3' and 806r; 5'-GGACTACHVGGGTWTCTAAT-3', annealing sequence) (Caporaso et al., 2012), and a single protist cell in a final reaction volume of 50 µl. PCR conditions were as follows: Initial denaturation was at 94°C for 3 minutes followed by 35 cycles of 94°C for 45 seconds, 50°C for 60 seconds, 72°C for 2 minutes. Final extension was at 72°C for 10 minutes (Benjamino and Graf, 2016). For *P. vertens* and *D. gracilis* primers 18SFU; 5'-ATGCTTGTCTCAAAGGRYTAAGCCATGC-3' and 18SRU; 5'-CWGGTTCACCWACGGAAACCTTGTTACG-3' were used (Tikhonenkov et al., 2016).

PCR products were separated and visualized using a 1% agarose gel, purified using either the Wizard® SV Gel and PCR Clean-up System (Promega) or the Monarch® DNA gel

Extraction kit (New England Biolabs® Inc.), and quantified using Qubit™ Flourometric quantitation (ThermoFisher Scientific). Barcoded V4 16S rRNA gene amplicons were pooled at 4nM equimolar concentrations and sequenced on an Illumina Miseq (Benjamino and Graf, 2016).

18S rRNA gene amplicons were cloned using the pGEM®-T Easy Vector System (Promega) following the manufacture's protocol and sequenced by Sanger sequencing. In addition to protist samples, negative controls consisting of TU, TE, and protist-free technical controls were amplified and sequenced. If needed, additional isolated protist cells were used in 18S rRNA gene-only PCR reactions and the amplicons were cloned and sequenced as described above.

V4 16S amplicon analysis using ASV inference

V4 16S rRNA gene reads were analyzed using a single workflow in R. Raw V4 amplicon reads were initially processed using the R package DADA2 (Callahan et al., 2016). Reads were quality trimmed, filtered, merged, and exact biological sequences (ASVs) were determined using the Divisive Amplicon Denoising Algorithm (DADA) (Callahan et al., 2016). This allows for determining the true biological sequence of V4 amplicons by applying the error rates observed within samples (Callahan et al., 2016, 2017). Taxonomy was then assigned to ASVs in DADA2 using the SILVA rRNA gene database Version 132 (Quast et al., 2013). Reagent contaminants were then identified and removed using the R package decontam (Davis et al., 2018) using the prevalence method with the threshold set to "0.35". This allows one to identify contaminating ASVs by assessing the prevalence of ASVs in negative control samples compared to true samples (Davis et al., 2018). Next, uncharacterized ASVs were removed from samples. Using taxonomic assignments, ASVs were then filtered based on their ecologically plausibility such

that only previously reported protist-associated bacterial taxa remained using the R package Phyloseq (McMurdie and Holmes, 2013). We then removed ASVs that were not found in at least three of the protist samples and at least 1 x 10^{-5} in their relative abundance. Exact filtering commands are provided within the amplicon workflow in Supplementary File 1. These filtering parameters are consistent with recent studies pertaining to identifying contaminants in low-biomass samples (de Goffau et al., 2018). Ordination analysis, α-diversity (Observed, Chao1, and Shannon), and β-diversity analysis (Bray-Curtis metric) were done using Phyloseq and Vegan (Oksanen et al., 2008) and the data were plotted using the R package ggplot2 (Wickham, 2011). For ordination analysis, a NMDS plot was generated using the Bray-Curtis metric and ellipses were generated for each protist genera per colony, representing the 95% confidence intervals. Statistical significance between α-diversity measurements were done using the Mann-Whitney U test in GraphPad Prism (version 8). Adonis which is part of the Vegan R package was used to perform PERMANOVA analyses. The entire V4 amplicon workflow is provided as a single R script which uses the raw reads generated by this study and reproduces all analyses and generates the resulting graphs (Supplementary File 1).

Phylogenetic analysis of SSU rRNA genes

ASVs corresponding to *Endomicrobium* were aligned to V4 16S rRNA gene reference sequences using MUSCLE (Edgar, 2004). Phylogenetic trees were generated using a Maximum likelihood (ML) method with the program IQ-TREE with model testing (Nguyen et al., 2015). The 18S rRNA genes obtained by this study were also aligned to reference sequences using MUSCLE and a phylogenetic tree was made using IQ-TREE with model testing.

Scanning electron microscopy

Scanning electron microscopy (SEM) was used to investigate the morphology of hindgut protists and their ectosymbionts. Protist cells were collected by low spin centrifugation as described above and fixed in 2% glutaraldehyde in TU (pH 7) for 1 hour at RT in an anaerobic chamber. The samples were deposited onto poly-L-lysine coated silicon wafer chips (Prod No. 16008, Ted Pella Inc.), washed with 80 mM Na cacodylate buffer (pH 7), and post-fixed in 2% osmium tetroxide at RT for 1 hour. The cells were rinsed twice for 5 minutes in distilled water then dehydrated in serial concentrations of ethanol (30%, 50%, 70%, 95%, 100%, 5 min each), and critical point dried (931GL, Tousimis). Samples were then mounted on SEM stubs using silver paint, sputter coated with palladium (E5100, Polaron), and examined using a scanning electron microscope (Nova NanoSEM 450, FEI).

Live fluorescent symbiont transmission assays

For all live-transmission assays, experiments were carried out in an anaerobic chamber with gas composition as described above. Hindguts were dissected from termites, ruptured with sterile forceps, and their contents were collected in anaerobic buffer containing anaerobic water with resazurin (1µg/ml), sodium thioglycolate (0.5g/L), and sodium bicarbonate (0.1M) pH 8.0 (Pedro et al., 2004). Samples were then fractionated by low spin centrifugation (3,000 rpm for 90 seconds) to separate protists and from bacteria which were unattached to protists. Each fraction was then washed three times in buffer by centrifugation at either 3,000 rpm (for protist fraction) or 13,000 rpm (for bacterial fraction) for 90 seconds. The washed fractions were then split into two equal volume groups and stained with either Texas Red®-X succinimidyl ester (TRSE, Molecular Probes™) or AlexaFlour 488 succinimidyl ester (SE488, Molecular Probes™) at 10µg/ml for 1 hour at room temperature (RT) in the dark in an anaerobic chamber. Covalent dye conjugation was done per manufacturer's instructions. Stained cells were then washed three

times in TU with reduced glutathione serving as a stop reagent for the amine reactive dyes. Protist and bacterial fractions were combined to produce two samples (Red-TRSE-stained and Green-SE-488 stained).

To assay for symbiont acquisition by protists, the two samples (Red-TRSE-stained and Green-SE-488 stained) were combined and monitored for the horizontal acquisition of new bacteria which was evident by heterogeneity in fluorescent signals of bacteria on individual protists. Samples were taken at various time points (0, 3, 15, and 20 hours), fixed with 2% formaldehyde, and viewed using a Nikon TE300 Eclipse microscope. Alternatively, fixed samples were mounted in ProLong™ Diamond Antifade Mountant (ThermoFisher) and imaged using a Nikon A1R Spectral Confocal microscope. To determine specificity of binding locations on protist hosts the positions of newly acquired ectosymbionts were counted along the length of two *Dinenympha* species (*D.* species II and *D. fimbriata*) and correlations in binding towards the anterior or posterior cell poles were tested using the Pearson R correlation test in GraphPad Prism (version 8). To test if symbiont acquisition required biologically active processes, this assay was repeated with the addition of either tetracycline 10µg/ml or cycloheximide 10µg/ml to each sample 1 hour prior to the start of the assay and compared to a no-treatment control. In addition, anaerobic symbionts were killed by exposure to atmospheric oxygen, labeled with propidium iodide (PI), and mixed with live cells to assay for the binding of dead bacteria to live protist hosts.

The fluorescent assay was then used to investigate whether ectosymbionts could come from the free-living (unattached) pool of bacteria. Hindgut contents were fractionated into bacterial or protist fractions and stained with TRSE as described above. These TRSE labeled fractions where then added to a SE-488 stained protist cell fractions and incubated in an

anaerobic chamber as described above. Samples were taken at 15 hours post the start of the assay, fixed, and viewed as described above.

Results

Morphological and phylogenetic diversity of hindgut protists

The morphology of protists used in this study was investigated using both light microscopy and SEM. These data along with their 18S rRNA gene phylogeny indicated that these protists consisted of six different species: *T. agilis*, *P. vertens*, *D. gracilis*, *D. fimbriata*, and two uncharacterized *Dinenympha* species (I & II). We obtained near-full length or partial (>1 kb) 18S rRNA genes sequences from individual protist cells, aligned them to references sequences, and reconstructed their phylogeny using IQ-TREE. Undescribed species such as *Dinenympha* species I & II clustered within other *Dinenympha* sequences supporting that they are indeed members of that genus (Figure 1). Differential interference contrast (DIC) micrographs of representative morphotypes of each protist species used in this study are provided as Supplementary Figure 2.

ASV composition of protist-associated bacteria

Our high-throughput amplicon analysis of the V4 region of the bacterial 16S rRNA gene recovered 102 bacterial ASVs from single protist cells spanning six different protist species. The most diverse bacterial taxon was Spirochaetes belonging to the genus *Treponema* which accounted for 66 of the 102 ASVs (Figure 2A). Other taxa included Bacteroidetes (17 ASVs), Elusimicrobia (*Endomicrobium*) (7 ASVs), Proteobacteria (7 ASVs), Verrucomicrobia (3 ASVs), Margulisbacteria (1 ASV), and Synergistetes (1 ASV) (Figure 2A).

To determine if there were any differences between the communities of bacterial symbionts within and across different protist species α-diversity and β-diversity analyses were conducted using the Phyloseq and Vegan R packages. Despite being smaller in size, *Dinenympha* species hosts were associated to a more diverse bacterial community compared to both *Trichonympha* (Mann-Whitney U test Observed, Chao1, and Shannon α-diversity $p < 0.0001$) and *Pyrsonympha* (Mann-Whitney U test Observed, Chao1, and Shannon α-diversity $p < 0.0001$) protists (Figure 2B). These protist-associated communities grouped together according to their protist hosts using the Bray-Curtis metric to perform β-diversity ordination (Figure 2C). These communities were significantly different from one another and were influenced by the genus of their host protist (PERMANOVA Bray-Curtis f=24.188, R^2=0.47253, p=0.001) and to lesser extent, the termite colony which they were isolated from (PERMANOVA Bray-Curtis f=1.8819, R^2=0.06516, p=0.041). The effect of termite colony was further investigated within each group of protist hosts. For *Trichonympha* hosts, the bacterial communities were heterogeneous with respect to termite colony (PERMANOVA Bray-Curtis f=3.3298, R^2=0.25954, p=0.019) which was also true for *Dinenympha* (PERMANOVA Bray-Curtis f=3.1712, R^2=0.25027, p=0.001) however, the bacterial communities of *Pyrsonympha* hosts were homogenous across different termite colonies (PERMANOVA Bray-Curtis f=0.633226, R^2=0.11226, p=0.677).

Several ASVs corresponding to ectosymbionts (Spirochaetes) were observed to be present across almost all of the *Dinenympha* species protist samples which included four different protist species (Figure 2A). In addition, these ASVs were also present (but at much lower prevalence and abundance) on some of the *Trichonympha* and *Pyrsonympha* protist cells.

These observations warranted further study and prompted us to investigate their mode of transmission (see below).

Phylogenetic diversity of protist-associated *Endomicrobium*

Since previous studies concluded that the *Endomicrobium* endosymbionts demonstrate co-speciation with their protist host as inferred by congruent phylogenies, we investigated which of the *Endomicrobium* (Elusimicrobia) ASVs demonstrated that same congruence. This was possible since we were able to co-amplify and sequence some of the 18S rRNA genes of protist cells used in our high-throughput amplicon study. This allowed us to determine which *Endomicrobium* ASVs were likely to be endosymbionts rather than contaminates from micromanipulation or from protist feeding activity, or transient symbionts. The results from these phylogenetic analyses suggested that five out of the seven *Endomicrobium* ASVs shared congruent evolutionary histories with their protist hosts while two ASVs did not (ASV6 and ASV7; Supplementary Figure 3). These two ASVs may represent contaminants or transient symbionts. The remaining five *Endomicrobium* ASVs clustered into three distinct groups, each associating with either *Trichonympha* (2 ASVs), *Pyrsonympha* (1 ASV), or *Dinenympha* (2 ASVs) (Supplementary Figure 3). The two *Endomicrobium* ASVs which were associated with *Trichonympha* hosts differed from one another by one base pair and did not appear to co-colonize individual *Trichonympha* cells. This trend was also true for the two *Endomicrobium* ASVs which associated with *Dinenympha* species II hosts. We did detect some instances in which some of the *Endomicrobium* sequences from one host were present in other protist samples. For example, one of the *Endomicrobium* ASVs from *Trichonympha* hosts were found in some *Pyrsonympha* and *Dinenympha* samples, but only at low abundances (Figure 2A).

Horizontal transmission of ectosymbionts

The ASV analysis of bacterial symbionts of single protist cells suggested that some ectosymbiont types (Spirochaetes) associated with multiple protist species. We hypothesized that these molecular data indicated symbiont sharing that may have come about through the horizontal transmission of ectosymbionts. An *in vitro* fluorescence-based assay was developed to test this hypothesis. Protists and bacteria from the hindgut of *R. flavipes* were covalently stained with either TRSE (red fluorescence) or SE488 (green fluorescence), mixed together, and the transfer or acquisition of new ectosymbionts was assayed over time. Since protists began the experiment with ectosymbiont communities that were homogeneous in their fluorescent label, newly acquired ectosymbionts were evident based on fluorescent heterogeneity of ectosymbiotic bacteria. Transmission observed in this assay should represent only half of the total transfer events since we could not distinguish newly acquired ectosymbionts which were the same color as the majority of the cells on the host. Over time, many species of protist hosts including *T. agilis* (Figure 3K – O) and several species of *Dinenympha* acquired horizontally transferred ectosymbionts (Figure 3), which were visibly attached to host's plasma membrane (Figures 3E, 3J, and 3O), not entangled in flagella or other bacterial cells, and were never observed in non-mixed control samples (Supplementary Figure 5). These data supported the hypothesis that some ectosymbionts of hindgut protist can be horizontally acquired.

In well characterized symbioses in which symbionts are horizontally transmitted, several active biological processes are involved. These include changes in the gene expression of the symbiont so that it can properly recognize and occupy its niche on or in its host (Bright and Bulgheresi, 2010; McFall-Ngai, 2014). To determine if ectosymbiont acquisition by hindgut protists requires active processes, we tested whether inhibiting protein synthesis would affect ectosymbiont transmission. The assay was repeated with the addition of either tetracycline or

cycloheximide and compared to a no-treatment control. Tetracycline, which inhibits translation, was chosen as the bacteriostatic agent due to previous reports that termite-associated *Spirochaetes* and *Bacteriodetes* decreased in their relative abundance after tetracycline treatment, suggesting that they were sensitive to that antibiotic (Peterson et al., 2015). Cycloheximide has been used to target protein synthesis across different protist taxa (Corno and Jürgens, 2006; Kodama and Fujishima, 2008) and was used in this study in an attempt to interfere with translation in the hindgut protists of *R. flavipes*. Over time, samples which were treated with tetracycline had significantly fewer protists that acquired new ectosymbionts compared to the no-treatment control (Two-tailed T-test, Time = 20 hours, $t = 5.278$, $df = 3$, $p = 0.0133$) (Figure 4A). These data indicated that inhibiting protein synthesis in the ectosymbionts affected their ability to be horizontally acquired by their protist hosts. Samples which were treated with cycloheximide were not significantly different from the no-treatment control.

In addition to inhibiting protein synthesis, we exposed samples consisting of both protists and bacteria to atmospheric oxygen for several hours, which killed strictly anaerobic organisms. We confirmed that oxygen killed both ectosymbiont and free-living bacteria by labeling with propidium iodide (PI) which labels cells which have died (Boulos et al., 1999) (Figure 4B and 4C). These PI-labeled cells were then added to live samples to assay for the binding of dead ectosymbionts to live protist hosts. In these experiments, we did not observe the binding of dead ectosymbionts to live protist cells ($n = 4$ independent experiments) (Figure 4D). We concluded from these experiments that the horizontal transmission of ectosymbionts requires live ectosymbionts and active translation. These data also support that the horizontal transmission observed in our assays is not due to non-specific binding.

We noticed that during these experiments, most newly acquired ectosymbionts appeared to bind to the anterior end of *Dinenympha* species II. To determine if this was true, or if binding was random, newly attached ectosymbionts were counted along the length of this protist species. The resulting data supported that newly acquired ectosymbionts bound more frequently towards the anterior cell pole of *D*. species II (Pearson's R p=0.0005) (Figure 4E-4I) than the posterior cell pole. This increase in frequency at one cell pole compared to the other was not observed in other *Dinenympha* species (Figure 4E). Since this cell pole is lacking flagella from the host cell, is it unlikely that this increased binding was an artifact due to entanglement. These data supported that the binding of ectosymbionts to protist hosts was not a random event, and that in *Dinenympha* species II there was a preferred region for the acquisition of new, horizontally acquired ectosymbionts.

Ectosymbionts in the free-living fraction

After observing the horizontal transmission of ectosymbionts we decided to sequence the bacterial V4 16S rRNA genes from the free-living bacterial fraction of three termite hindgut samples to see if we could detect protist-associated *Treponema* ASVs in those fractions. In each of the three free-living bacterial fractions we detected protist-associated *Treponema* ASVs which on average accounted for 72.7% of the total *Treponema* reads generated from the free-living bacterial fractions (Supplemental Figure 4A). These data showed that some *Treponema* that are normally found attached to protists were also present in these hindguts unattached to protist hosts. However, we cannot rule out that this may have been influenced, at least in part, to the detachment of some ectosymbionts from their protist hosts during sample preparations.

After detecting ectosymbiont *Treponema* ASVs in the free-living bacterial fractions of hindguts, we used our fluorescence assay to determine if newly attached ectosymbionts could

also transfer to protists from the pool of free-living bacteria. In this assay, horizontal transmission was seen from ectosymbionts from both the free-living bacterial fraction containing unattached bacterial cells, as well as from the protist cell fractions, where bacteria were mainly attached to protists (Supplementary Figure 4B). There was no significant difference between the percentage of protist cells that acquired new ectosymbionts from the free-living or protist-associated fractions (Two-tailed T-test, $t = 1.054$, $df = 2$, $p = 0.402$).

Discussion

In this study, we show that in the termite *R. flavipes*, the associations between hindgut protists and their symbiotic bacteria exhibit specificity in different aspects of their interactions including (i) host range, (ii) community structures and (iii) transmission trends. We also show that bacterial communities associated with certain genera of protist hosts vary with respect to diversity (Figure 2B). *Dinenympha* hosts associated with a more diverse community of bacteria compared to larger protist species such as *P. vertens* and *T. agilis* (Figure 2B). Since most of the ASVs which associated with *Dinenympha* were *Treponema* (Figure 1A), these data suggested that the ectosymbiont communities on single *Dinenympha* cells were more diverse compared to others. These ectosymbiont communities also included ASVs which were shared across many of the *Dinenympha* species and led to the hypothesis that at least some ectosymbionts can be horizontally acquired. This was experimentally supported using a fluorescence-based assay, which allowed us to visualize the acquisition of new ectosymbionts by protist cells over time, and test whether acquisition required biologically active processes.

In general, the bacterial communities of these protist species were distinct from one another (Figure 2C) as most bacterial ASVs were exclusive to certain protist hosts (Figure 2A). Despite occurring in the same hindgut, associations between protist cells and their symbionts

exhibited varying levels of fidelity. For example, even though many ectosymbiont ASVs were shared across closely related *Dinenympha* species, only a few were putatively shared with *Trichonympha* or *Pyrsonympha* (Figure 2A). This suggested that there must be mechanisms that result in, or ensure, specificity between these bacterial symbionts and a specific range of related protist hosts. The increase in *Treponema* diversity associated with *Dinenympha* hosts may be due to common features such as attachment factors which are shared among protist hosts who are more closely related and co-inhabit the same termite hindgut. These protist-symbiont associations were also fairly conserved across different termite colonies with only a small amount of the variation within these groups of protists being due to a colony effect (Figure 2C).

The ASV analysis provided a high-resolution characterization of the community structures and transmission trends of the bacterial symbionts of hindgut protists. For example, individual *T. agilis* cells differed with respect to which *Endomicrobium* ASV they were associated. For example, two *T. agilis* cells (cells A and B) which shared the same *Endomicrobium* ASV were more similar to one another in their 18S rRNA gene sequence than they were to a third cell (cell C, Figure 1A) which associated with a second *Endomicrobium* ASV. These two *Endomicrobium* ASVs differed by just one base pair in the V4 region of their 16S rRNA gene. We have since obtained genomes for each of these *Endomicrobium* ASVs and genome analysis supports that they are not clonal but likely two distinct genomovars or putative strains (unpublished data). These differences in the symbiont communities of related protists may reflect their evolutionary divergence from one another. Previous studies have already demonstrated the possibility that what was thought to be a single species of *T. agilis* in *R. flavipes* is likely more than one species (James et al., 2013).

Overall, the community structures of these protist-associated bacteria differed from one another with respect to the ecological niche that they occupied on or within their protist hosts. For example, ectosymbiotic bacteria such as the *Treponema* (66 ASVs) and '*Ca*. Symbiothrix' (6 ASVs) presented more sequence diversity compared to the intracellular *Endomicrobium* (7 ASVs) or *Verrucomicrobia* (3 ASVs) symbionts. This may be due to different ecological factors encountered across the intracellular or extracellular niches. For example, the extracellular environment is likely to be dynamic compared to the cytoplasm of a protist host. Ectosymbionts may experience differences in available nutrients and be subject to greater competition compared to endosymbionts. Supporting this, the ectosymbiont '*Ca*. Symbiothrix dinenymphae', does not show evidence of genome reduction and encodes many genes evolved in polysaccharide degradation, and the uptake of various sugar monomers (Yuki et al., 2015) suggesting that these ectosymbionts may need the genomic and metabolic flexibility to use different carbon sources when available. Such selective pressures may be responsible for the increased diversity seen in protist-associated ectosymbionts communities compared to the communities of endosymbiotic bacteria. It is also plausible that the extracellular niche of these protists may be more easily colonized by bacteria compared to their intracellular niche.

In lower termites, protist-associated *Treponema* are diverse with members belonging to at least two phylogenetic clusters (Iida et al., 2000; Ohkuma, 2008). The divergence of these *Treponema* clusters is not due to the phylogeny of their termite, or protist, hosts as both clusters contain *Treponema* found associated with various termites and protists species. Our data supports that these ectosymbiotic communities of *Treponema* are diverse and that *Dinenympha* hosts associated with more diverse *Treponema* communities compared to other protists.

The associations between protists and the *Treponema* ectosymbionts are thought to involve metabolic interactions. It has been hypothesized that symbiosis between hindgut protists and their ectosymbiotic *Treponema* involves syntrophic exchange of reduced fermentation end products such as CO_2 and H_2. This hypothesis stems from the observation that cultivated strains of *Treponema primitia*, belonging to the termite *Treponema* phylogenetic cluster I, have been shown to consume CO_2 and H_2 in reductive acetogenesis, as well as fix nitrogen (Graber et al., 2004; Graber and Breznak, 2004; Leadbetter et al., 1999; Noda et al., 2003). In addition, analysis of expression profiles for formyl tetrahydrofolate synthetase, a key enzyme in the reductive acetogenesis pathway, supported that *Treponema* are the primary acetogens in the hindgut of *R. flavipes* and other termite species (Pester and Brune, 2006). Recently, a draft genome of an ectosymbiotic *Treponema* sp. (NkOx-clu11) of a termite-associated *Oxymonas* sp. was shown to encode genes involved in the Wood-Ljungdahl pathway (Utami et al., 2019). If other protists-associated *Treponema* are also acetogens, then they could provide their host with a necessary H_2 and CO_2 sink and may even fix nitrogen which is otherwise limiting in the termite's hindgut.

The ectosymbiont communities of hindgut protists in *R. flavipes* we studied were also dynamic as demonstrated by the fact that protists acquired new ectosymbionts over time. This horizontal transmission of ectosymbionts required active bacterial processes because it was lowered in the presence of tetracycline (Figure 4A). However, it could not be determined if inhibition of bacterial translation by tetracycline was directly involved in inhibiting the transmission. For example, tetracycline may have inhibited translation of proteins involved in host binding, or the effect could be due to a decrease in proteins involved in motility, other cellular processes, or an increased in bacterial mortality. It is also possible that tetracycline could indirectly affect the protist host cells as a previous study has noted a decrease in protist cell

counts following exposure to this antibiotic (Peterson et al., 2015). To further test that host-binding is an active process and not passive, oxygen-killed hindgut bacteria were stained with PI and added to live samples (Figure 4B-4D). In these experiments, dead bacteria were never observed to be attached to live protist hosts indicating that the horizontal acquisition seen in our experiments is active and not due to non-specific binding or entanglement.

That ectosymbionts bound preferentially to the anterior cell pole of *D*. species II suggested that there was spatial specificity to the process (Figure 4E-4I). This spatial specificity was not observed on other *Dinenympha* host cells. The cause of this specificity is not known, but may be the result of new cell membrane, or binding structures for ectosymbionts, being added to the host at the anterior pole. Since these protist cells are morphologically polarized, it could also be that there are protist functions or signals that are specific to that cell pole that allowed ectosymbionts to bind more readily than at the other cell pole. This suggests that following binding at the anterior end, ectosymbionts may be transported to other locations on the protist cell. However, spatial specificity of ectosymbiotic *Treponema* on the surface of termite protists has been observed in a previous study using phylotype specific fluorescent in situ hybridization (FISH) probes (Noda et al., 2003).

Protists from *R. flavipes*, and other lower termites cannot yet be cultured, and this results in some limitations in the ectosymbiont acquisition assays. After 20 hours, most protist cells have died and lysed during the *in vitro* experiments. This limited the time over which the assay could be conducted. Because of this, we could not determine if ectosymbionts could also be vertically transmitted during protist cell division. For vertical transmission, ectosymbionts would have to remain attached to a dividing host cell such that the daughter cells directly inherit the ectosymbionts of the parent cell. We have not yet witnessed actively dividing hindgut protists

but there is no evidence to suggest that they would have to shed their ectosymbionts prior to, or during, cell division. We also could not discern if the acquisition of ectosymbionts required active protein synthesis by the protist hosts. We found that cycloheximide did not significantly affect ectosymbiont acquisition. However, it may still be the case that protein synthesis by the protist host is required for acquisition. For example, it could be that the rate of protein turnover in these protists is slow and that 20 hours was not enough time to detect an effect. It is also possible that these protists are not sensitive enough to cycloheximide for it to completely inhibit protein synthesis.

The observations that we could (i) detect protist-associated *Treponema* ASVs in samples enriched in free-living bacteria and (ii) detect horizontal transmission of new ectosymbiont from the free-living bacterial cell fractions, supports that these ectosymbionts may live both in the free-living state and as protist-bound ectosymbionts. Since not all the protist-associated ectosymbiont ASVs were detected in the free-living bacterial fractions we cannot rule out the possibility that some ectosymbionts are obligate symbionts and perhaps, vertically acquired. In addition, because these ectosymbiont communities are complex, we cannot rule out that possibility that some members are transient, and that the horizontal transmission observed in our assays is not due to those members binding to protists. We did not detect any reads belonging to '*Ca*. Symbiothrix' in two of the three free-living bacterial fractions of individual termites and reads pertaining to that taxon were only present at 0.3% relative abundance in the third sample. These may represent ectosymbionts which are obligate or vertically acquired. Using FISH with ASV specific probes may help to resolve which symbionts are horizontal transmitted. However, because as many as 30 or more *Treponema* ASVs associated with a single protist cell (as is the

case with some *D*. species II cells), designing and testing that many FISH probes would be impractical.

The use of single protist cells as templates for high-throughput amplicon sequencing of the V4 region of their bacterial symbiont's 16S rRNA gene combined with the high-resolution of the DADA2 analysis, provided a detailed survey of these bacterial communities. This allowed us to investigate similarities and differences both within certain groups of protists such as the *Dinenympha*, but also across more distantly related protists such as *Trichonympha* and *Pyrsonympha* which co-inhabit the same hindgut. Although they associated with bacteria belonging to the same taxa, such was the case for *Treponema* and *Endomicrobium,* the communities of these different protists were distinct from one another and relatively conserved across termite colonies. It is important to note that although we carefully isolated these single cells and washed them by micromanipulation, it is still plausible that either bacteria or DNA were carried over during isolation. Since our data supports that these communities were distinct from one another we do not suspect that contaminates not identified by our filtering methods to have affected our analyses. We did however detect some minor instances of cross contamination of some *Endomicrobium* ASVs across different protist species which were at a low prevalence and low abundance. Increasing the number of washing steps or more conservative filtering criteria may reduce or eliminate these issues in future studies.

The assay to detect horizontal transmission of ectosymbionts was useful for revealing new information about the interactions between bacteria and protists in the termite hindgut community. The possibility of horizontal transmission of ectosymbionts between protists or between the pool of free-living bacteria and protists may explain how these ectosymbiont communities maintain their heterogeneous colonization of protists. Also, it may provide

information as to how some ectosymbiont types associate with high specificity; these may not participate in horizontal transfer and may, instead, be acquired vertically. Of course, specificity can occur even if bacteria are horizontally acquired. Specificity during transfer must result from mechanisms that encourage the binding of some symbionts and/or discourage the binding of others, perhaps through specific attachment factors made the host, the symbiont or by both. In some described cases of horizontal transmission, these attachments factors include bacterial pili or other proteins which bind to glycosylated surface proteins on the symbiont's host cells (Bright and Bulgheresi, 2010). Specificity may also involve chemotaxis, which could allow some ectosymbionts to find their respective protist hosts in the termite's hindgut. These attractants may include metabolites such as short chain fatty acids produced by the protist hosts as fermentation end products. Chemotaxis during horizontal transmission has been observed in other symbioses such as the case with *Vibrio fischeri*, which is attracted to chitobiose produced by its hosts, the Hawaiian bobtail squid (*Euprymna scolopes*) (McFall-Ngai, 2014). Chemotaxis has also been shown to play an important role in the pathogenesis of certain spirochetes such as *Treponema denticola* which rely on chemotaxis to navigate to certain host tissues (Lux et al., 2001). Taken together, chemotaxis and specific attachment factors may provide horizontally acquired ectosymbionts the ability to colonize their protist hosts with high fidelity and may enable the hypothesized metabolite exchange.

Future studies could investigate the use of different combinations of universal primers sets to increase the success of co-amplifying the 18S rRNA genes from protist hosts along with the V4 (or other regions) of the bacterial 16S rRNA genes. Another way in which these single cell assays could be explored or approved upon, would be to investigate the use of an 18S rRNA primer set which could amplify a variable region that would allow for high-throughput amplicon

sequencing. This would remove the need to clone and sequence 18S amplicons by sanger sequencing and likely yield higher coverage at a reduced cost. The use of high-throughput amplicon sequencing to investigate protist diversity in termite hosts is already in use (Jasso-Selles et al., 2017).

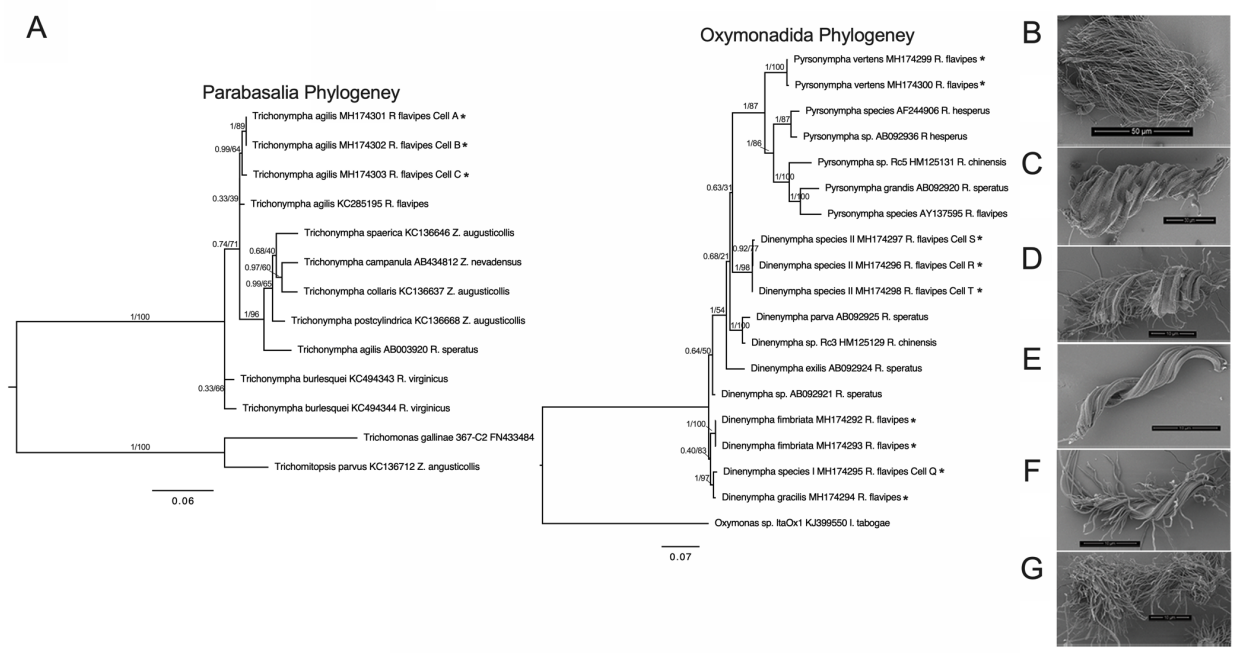


Figure 1 Phylogenetic and morphological diversity of hindgut protist species from *R. flavipes*. (A) Maximum likelihood (ML) phylogenetic trees of 18S rRNA genes from single protists cells. Parabasalia ML tree (left) was generated using the program IQ-Tree with substitution model GTR+I+G4 while the Oxymonadida ML tree (right) was generated using the substitution model TN+I+G4. Four 18S rRNA genes from single protist clustered to known references sequences (*D. fimbriata*, *D. gracilis*, *P. vertens*, and *T. agilis*). Other previously undescribed protists, *D*. species I and *D*. species II, clustered within the genus *Dinenympha*. Taxa marked by (*) represents sequences obtained by this study. Sequences from cells which are designated by a letter represent individuals in which the 18S rRNA gene was co-amplified with the bacterial V4 16S rRNA gene. Tip labels include the protist name, accession number, and host termite species. Scanning electron micrographs of representative individuals of each protist species ((B) *T. agilis*, (C) *P. vertens*, (D) *D*. species I, (E) *D. gracilis*, (F) *D. fimbriata*, and (G) *D*. species II). Scale bars represent 50µm (C), 30µm (D), and 10µm (E – G). Support values represent the Bayesian posterior probability and Bootstrap support values respectively.

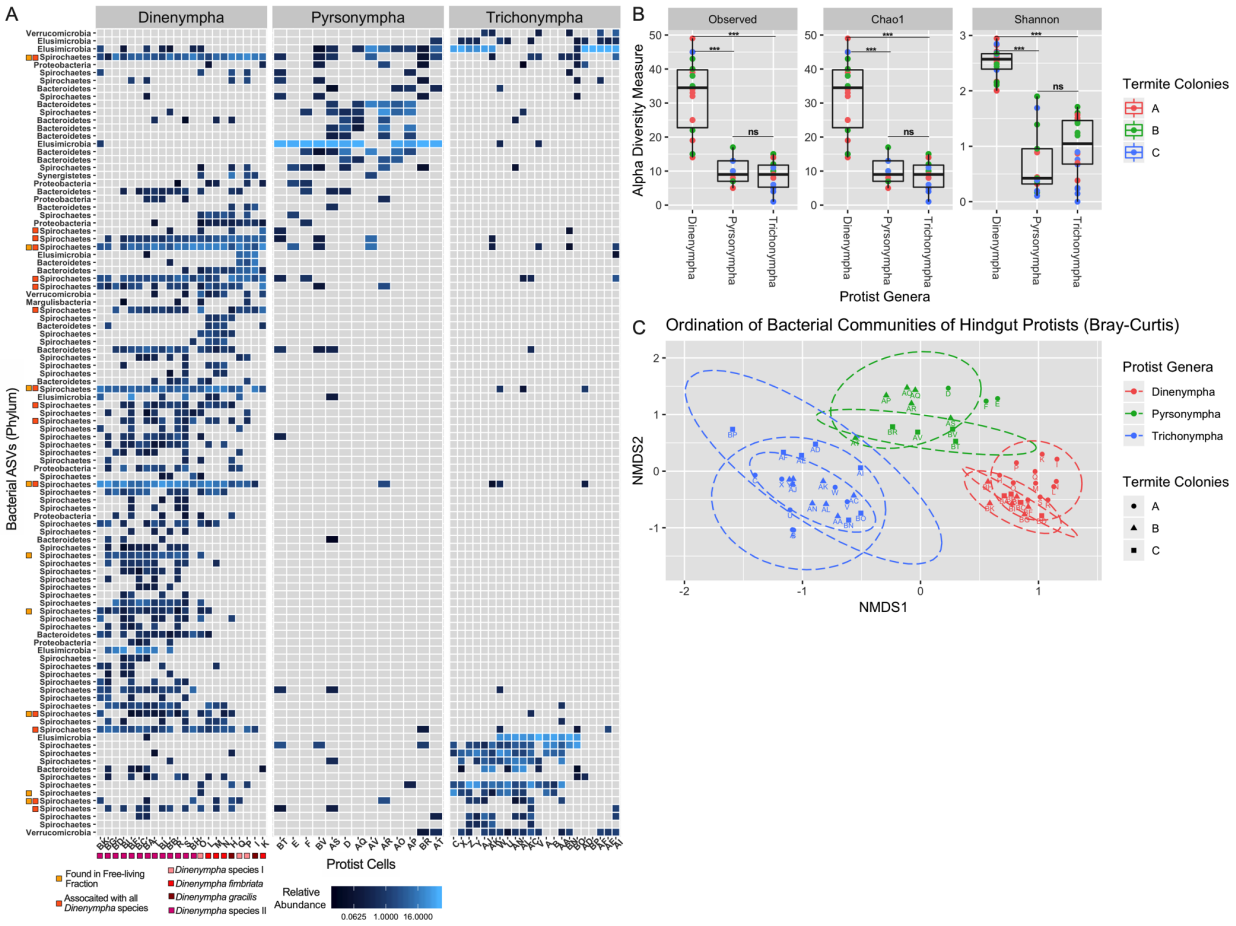


Figure 2 Diversity and distribution of protist-associated bacterial ASVs. (A) Heat map showing the distribution and relative abundance of bacterial ASVs (rows) which are reported at the Phylum level of taxonomy, across all protist cells (columns). (B) α-diversity measurements (Observed, Choa1, and Shannon) of each of the three protist genera (*Dinenympha*, *Pyrsonympha*, and *Trichonympha*). Box plots represent the 25th and 75th percentiles with the line representing the median and whiskers representing the largest and smallest values, respectively. Statistically significance was calculated using Mann-Whitney U test (***, p< 0.0001, ns= no significant difference). (C) Bray-Curtis β-diversity ordination plot of the bacterial ASVs of protist cells. Each point represents the bacterial community of a single protist cell where the color represents the protist genus, shape indicates the termite colony, and the dashed ellipses represents the 95% confidence intervals for each protist genus per termite colony. Points are labeled by a letter which represents protist cell's identification.

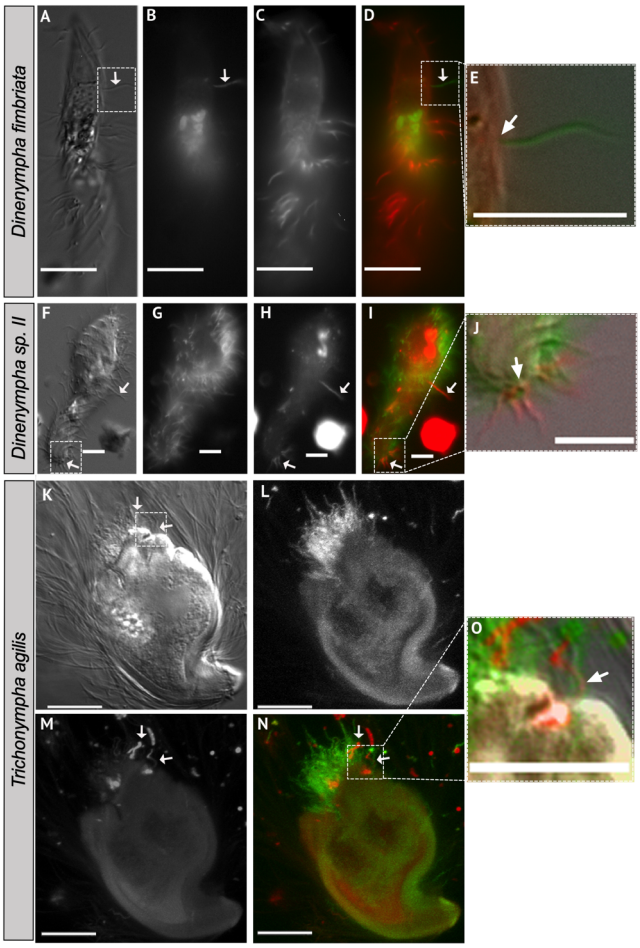


Figure 3 Horizontal transmission of ectosymbionts across different protist species. DIC and fluorescence micrographs of hindgut protist and their ectosymbionts stained with either TRSE (shown red) or SE-488 (shown green) at Time=12 hours of fluorescent assay. Overtime several different protist species including (A-O) *D. fimbriata*, (F - J) *D*. species II, and (K – O) *T. agilis* acquired new ectosymbionts. Micrographs are arranged from left to right as DIC, SE488, TRSE, and merged (SE488 and TRSE) for each protist. Fluorescence micrographs L – M are maximum intensity Z-projections. Micrographs E, J, and O represent areas which are zoomed in to demonstrate that the ectosymbionts are visible attached to the host cell's plasma membrane. Arrows point to horizontally acquired ectosymbionts and scale bars represent 10µm.

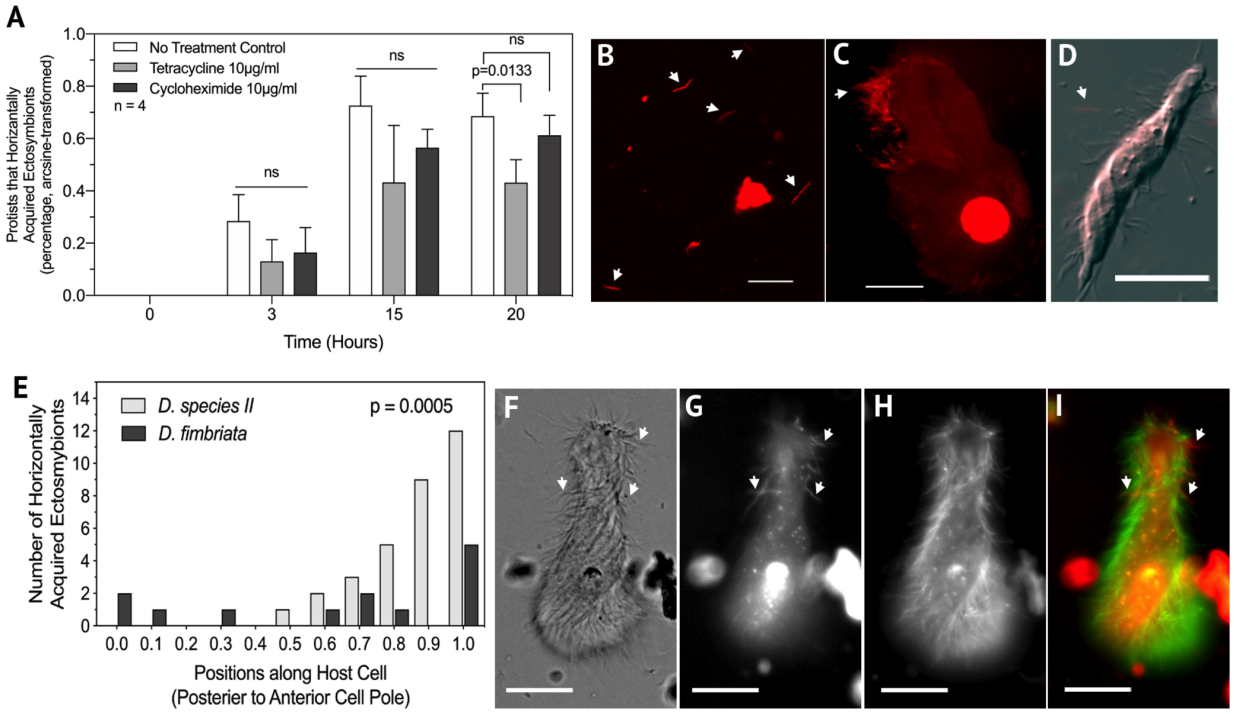
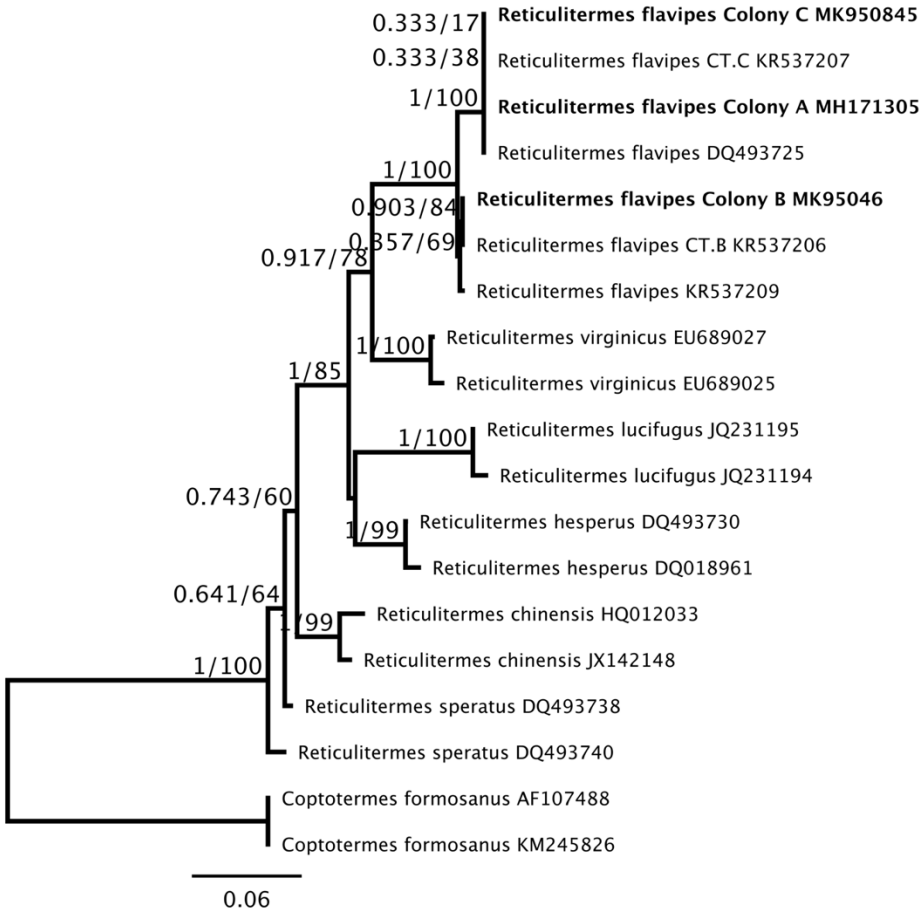
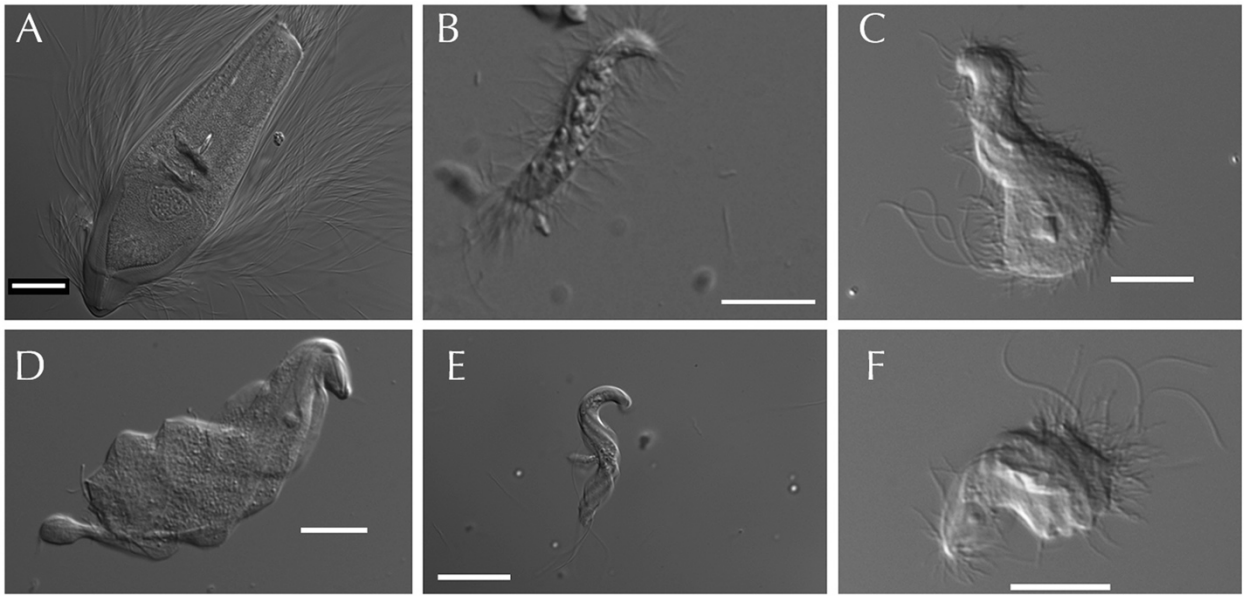


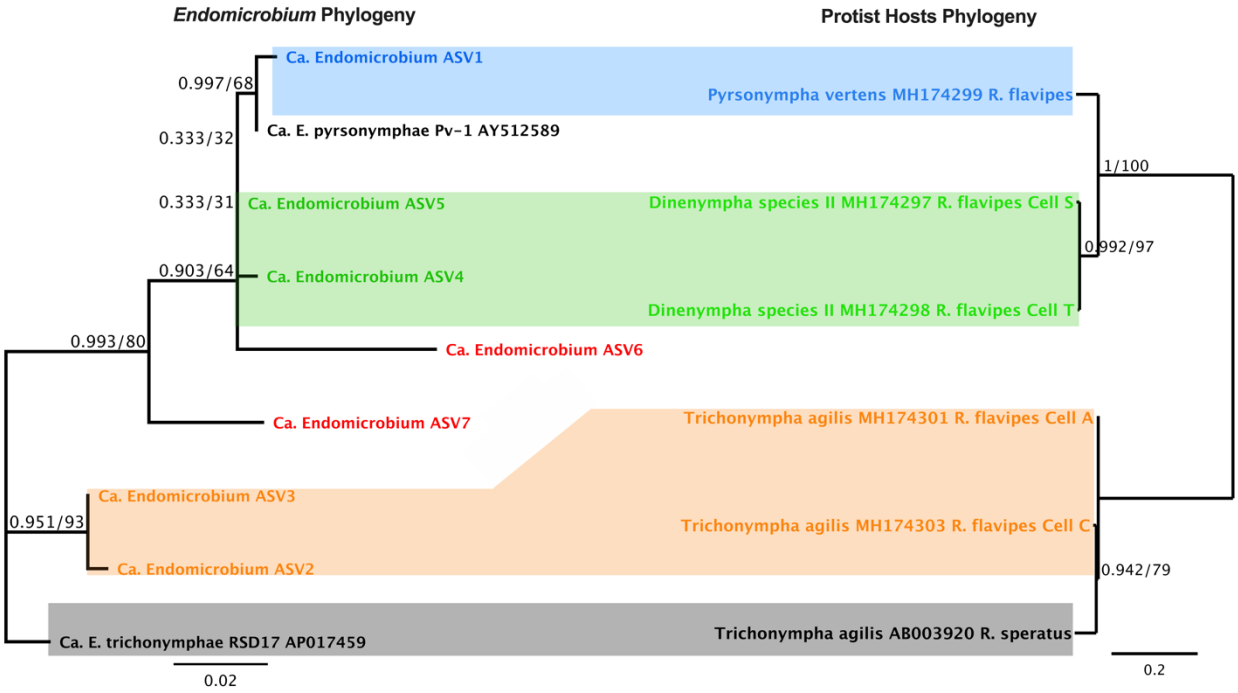
Figure 4 Horizontal transmission of ectosymbionts involves active processes and is non-random.(A) The addition of tetracycline significantly lowered the percentage of protists that acquired at least one new ectosymbiont (Two-tailed T-test, Time = 20 hours, t = 5.278, df = 3, p = 0.0133, n = 4 independent experiments) while the addition of cycloheximide had no significant effect. (B - D**)** Micrographs of PI stained cells. Exposing hindgut contents to O_2 killed hindgut bacteria (B and C) which did not bind to live protist cells (D) (arrows point to O_2 killed bacteria). (E) Significantly more ectosymbionts (Pearson's R, p=0.0005) bound towards the anterior cell pole compared to the posterior cell pole on *D*. species II however, this binding characteristic was not seen in other *Dinenympha* species. (F - I) Fluorescence and DIC micrograph of *D*. species II stained with amine reactive dyes (G TRSE, H SE488), showing increased binding of new ectosymbionts (arrows) toward the anterior cell pole. Scale bars represent 10µm.



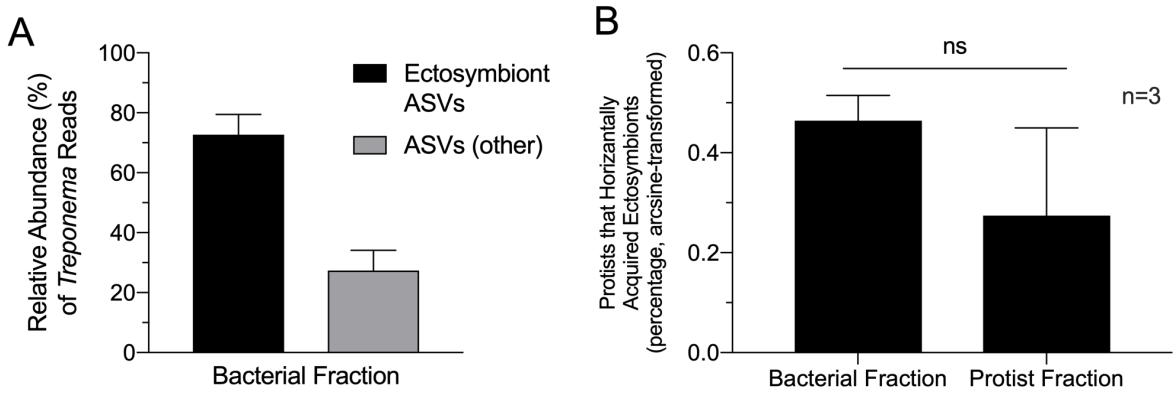
Supplementary Figure 1 Phylogenetic tree of termite mitochondrial cytochrome oxidase II gene sequences. Sequences were aligned to references with MUSCLE and a Maximum likelihood (ML) phylogenetic tree was made using IQ-Tree with substitution model TIM2+G4. Sequences obtained from termites used in this study (Bold) clustered within the *R. flavipes* clade. Support values represent the Bayesian posterior probability and Bootstrap support values respectively.



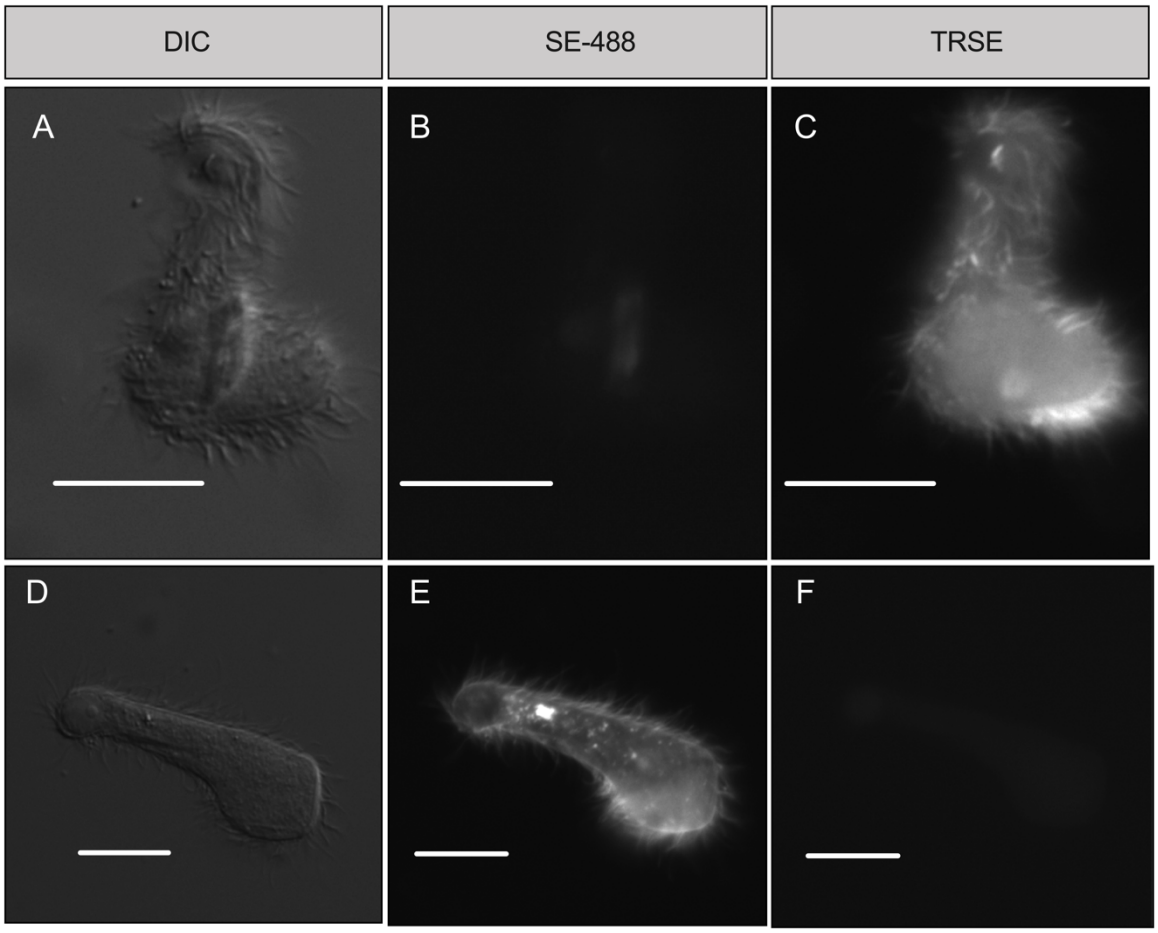
Supplementary Figure 2 DIC micrographs of representative morphotypes of each protist species used in this study. (A) *Trichonympha agilis*, (B) *Dinenympha fimbriata*, (C) *Dinenympha* species II, (D) *Pyrsonympha vertens*, (E) *Dinenympha gracilis*, and (F) *Dinenympha* species I. Although *D*. species I and *D*. species II are morphologically similar under light microscopy, some notable differences include the colonization patterns of their ectosymbionts, and the larger size of *D*. species II compared to *D*. species I. Scale bars represent 20μm.



Supplementary Figure 3 Congruent phylogenies between endosymbiotic *Endomicrobium* ASVs and their protist hosts. *Endomicrobium* ASVs were aligned to the V4 regions of the 16S rRNA sequence from relatives using MUSCLE and a Maximum likelihood (ML) phylogenetic tree was made using IQ-Tree with substitution model K2P+I. The 18S rRNA genes from single protist hosts were also aligned to themselves and one reference sequence using MUSCLE and a ML phylogenetic tree was made using IQ-Tree with substitution model TN+G4. Sequences are color coded such endosymbionts are colored the same as their host and putative transient ASVs or contaminates are colored Red. Support values represent the Bayesian posterior probability and Bootstrap support values respectively.



Supplementary Figure 4 Ectosymbionts in the free-living (unattached) bacterial fraction. (A) The relative abundance of ectosymbiotic *Treponema* ASVs out of the total *Treponema* reads sequenced from the bacterial fractions of individual termite hindguts (n=3). Bars represent the average relative abundance of each ASV type and error bars represent the standard error of the mean. (B) The average percent (arcsine transformed) of protists that acquired new ectosymbionts from either the protist or bacterial cell fractions of termite hindguts at Time=15 hours of the assays. There was no significant difference (Two-tailed T-test, $t = 1.054$, $df = 2$, $p = 0.402$) between these two groups. Error bars represent the standard error of the mean and n= 3 independent experiments.



Supplemental Figure 5 Non-mixed controls for fluorescent assays. Representative micrographs of protist cells from non-mixed control samples from fluorescent assays. Micrographs represent DIC (A and D) SE-488 fluorescence (B and E) and TRSE fluorescence (C and D). Control samples always maintained homogenous fluorescence. Micrographs were taken post 12 hours from the start of assays and scale bars represent 20μm.

Supplementary File 1. Custom R workflow used to filter and analyze V4 16S rRNA gene amplicons generated by this study.

```
# This R Script was pieced together by: Michael E. Stephens, University of
Connecticut, Storrs CT, USA
# Used to analyze bacterial ASVs associated with protists isolated from
the hindgut of the termite, Reticulitermes flavipes
# This script is provided for the reproducibility of all ASV-related
analyses used in this study
#
# This Pipeline uses raw R1 and R2 read files (.fastq format) for 16S rRNA
gene (V4 region) amplicons to generate ASVs, decontaminate (if controls
were used), filter, and generate graphs
# dada2 pipeline taken follows tutorial:
https://benjjneb.github.io/dada2/tutorial.html
# Other tutorials and references are provided throughout this R script;
they are all publicly available

# USER MUST SET PATHS TO APPROPIATE DIRECTORIES FOR READS, TAXONOMY FILES,
MAPPING FILE, AND ALL OUTPUT FILES

# Step 1
########################################################################
###########
# Load all packages
########################################################################

library(dada2); packageVersion("dada2")
library(phyloseq); packageVersion("phyloseq")
library(ggplot2); packageVersion("ggplot2")
library(vegan)
library(decontam)

# Packages for some downstream analyses, following:
https://www.bioconductor.org/help/course-
materials/2017/BioC2017/Day1/Workshops/Microbiome/MicrobiomeWorkflowII.htm
l

library("knitr")
library("BiocStyle")
.cran_packages <- c("ggplot2", "gridExtra")
.bioc_packages <- c("dada2", "phyloseq", "DECIPHER", "phangorn")
.inst <- .cran_packages %in% installed.packages()
if(any(!.inst)) {
  install.packages(.cran_packages[!.inst])
}
.inst <- .bioc_packages %in% installed.packages()
if(any(!.inst)) {
  source("http://bioconductor.org/biocLite.R")
  biocLite(.bioc_packages[!.inst], ask = F)
}
sapply(c(.cran_packages, .bioc_packages), require, character.only = TRUE)
```

```
# Step 2
##########################################################################
##########
#Load in Read files and visualize quality
##########################################################################
########################

# Set Path to direcotry with un-zipped read files (.fastq)

path <-""
list.files(path)

# Forward and reverse fastq filenames have format: SAMPLENAME_R1_001.fastq
and SAMPLENAME_R2_001.fastq

fnFs <- sort(list.files(path, pattern="_R1_001.fastq", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_R2_001.fastq", full.names = TRUE))

# Extract sample names, assuming filenames have format:
SAMPLENAME_XXX.fastq

sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)

# Visulaize quality profiles of Forward Reads

plotQualityProfile(fnFs[1:82]) # Set to the number of reads you want to
see

## Scale for 'y' is already present. Adding another scale for 'y', which
will replace the existing scale.

# Visulaize quality profiles of Reverse Reads

plotQualityProfile(fnRs[1:82]) # Set to the number of reads you want to
see


#Step 3
##########################################################################
###############
#Filter reads based on size, Trim, learn error rates, Dereplicate, and
Merge paired Reads ######################################################

# Place filtered files in filtered/ subdirectory

filtFs <- file.path(path, "filtered", paste0(sample.names,
"_F_filt.fastq.gz"))
filtRs <- file.path(path, "filtered", paste0(sample.names,
"_R_filt.fastq.gz"))

# Filter and Trim

out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(240,160),
```

```
                     maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
                     compress=TRUE, multithread=TRUE) # On Windows set
multithread=FALSE
head(out)

# Learn Forward and Reverse Error Rates

errF <- learnErrors(filtFs, multithread=TRUE)
errR <- learnErrors(filtRs, multithread=TRUE)

# Visualize Forward Error Rates

plotErrors(errF, nominalQ=TRUE)+
  ggsave("forward_errors.png", path = "")

plotErrors(errR, nominalQ =TRUE)+
  ggsave("reverse_errors.png", path = "")

# Derepilication
# Dereplication combines all identical sequencing reads into into
,Äúunique sequences,Äù with a corresponding ,Äúabundance,Äù equal to the
number of reads with that unique sequence.
# Dereplication substantially reduces computation time by eliminating
redundant comparisons.

derepFs <- derepFastq(filtFs, verbose=TRUE)
derepRs <- derepFastq(filtRs, verbose=TRUE)

# Name the derep-class objects by the sample names

names(derepFs) <- sample.names
names(derepRs) <- sample.names

# Apply the core sample inference algorithm (dada) to the dereplicated
data

dadaFs <- dada(derepFs, err=errF, multithread=TRUE)
dadaRs <- dada(derepRs, err=errR, multithread=TRUE)

# Merge Paired Reads
mergers <- mergePairs(dadaFs, derepFs, dadaRs, derepRs, verbose=TRUE)

# Inspect the merger data.frame from the first sample
head(mergers[[1]])

# Construct Sequence Table
seqtab <- makeSequenceTable(mergers)
dim(seqtab)
seqtab2 <- seqtab[,nchar(colnames(seqtab)) %in% seq(250,256)]

# Inspect distribution of sequence lengths
table(nchar(getSequences(seqtab)))
table(nchar(getSequences(seqtab2)))
```

```
# Step 4
##########################################################################
##########
# Remove Chimeras
##########################################################################
##################

# Remove Chimeras
seqtab.nochim <- removeBimeraDenovo(seqtab2, method="consensus",
multithread=TRUE, verbose=TRUE)
dim(seqtab.nochim)

# % of Reads Remaining
sum(seqtab.nochim)/sum(seqtab2)

# Track Reads Through the Pipeline so far
getN <- function(x) sum(getUniques(x))
track <- cbind(out, sapply(dadaFs, getN), sapply(dadaRs, getN),
sapply(mergers, getN), rowSums(seqtab.nochim))

# If processing a single sample, remove the sapply calls: e.g. replace
sapply(dadaFs, getN) with getN(dadaFs)
colnames(track) <- c("input", "filtered", "denoisedF", "denoisedR",
"merged", "nonchim")
rownames(track) <- sample.names
(track)

# Step 5
##########################################################################
################
# Assign Taxonomy, generate initial ASV counts and taxonomy tables
##########################################################################
#################

# Assign Taxonomy

taxa <- assignTaxonomy(seqtab.nochim, "silva_nr_v132_train_set.fa",
multithread=TRUE) #Path to taxonomy training set
taxa <- addSpecies(taxa,"silva_species_assignment_v132.fa") #Path to
second (species) taxonomy training set

# Inspect Taxonomy

taxa.print <- taxa # Removing sequence rownames for display only
rownames(taxa.print) <- NULL
head(taxa.print)

# Step 6
##########################################################################
################
# Align seqs, make a phylogenetic tree, and combine information into a
phyloseq object ############################################################
```

```
# Follows tutoiral found here: https://www.bioconductor.org/help/course-
materials/2017/BioC2017/Day1/Workshops/Microbiome/MicrobiomeWorkflowII.htm
l
# This can take a few minutes..or hours...

seqs <- getSequences(seqtab.nochim)
names(seqs) <- seqs # This propagates to the tip labels of the tree
alignment <- AlignSeqs(DNAStringSet(seqs), anchor=NA,verbose=FALSE)

phangAlign <- phyDat(as(alignment, "matrix"), type="DNA")
dm <- dist.ml(phangAlign)
treeNJ <- NJ(dm) # Note, tip order != sequence order
fit = pml(treeNJ, data=phangAlign)
fitGTR <- update(fit, k=4, inv=0.2)
fitGTR <- optim.pml(fitGTR, model="GTR", optInv=TRUE, optGamma=TRUE,
                    rearrangement = "stochastic", control =
pml.control(trace = 0))
detach("package:phangorn", unload=TRUE)

# Combine data into a Phyloseq Object
# Fix or confirm row names are sample name

samdf <- read.csv("")  # samdf = mapping file (.csv)
all(rownames(seqtab.nochim) %in% samdf$SampleID) # TRUE
rownames(samdf) <- samdf$SampleID
keep.cols <- c("sample_or_control", "Host_species", "Host_genus",
"Host_phyla", "cell_id", "sample_collection_date", "termite_id",
"termite_colony", "lab_colony_id", "colony_collection_date")
samdf <- samdf[rownames(seqtab.nochim), keep.cols]

#Make Phyloseq Object (ps)

ps <- phyloseq(otu_table(seqtab.nochim, taxa_are_rows=FALSE),
               sample_data(samdf),
               tax_table(taxa),phy_tree(fitGTR$tree))
ps

# Inspect ASV table #

# Extract abundance matrix from the unfiltered phyloseq object
OTU1 = as(otu_table(ps), "matrix")
# transpose if necessary
if(taxa_are_rows(ps)){OTU1 <- t(OTU1)}
# Coerce to data.frame
OTUdf = as.data.frame(OTU1)
write.table(OTUdf, file = "", sep = " ")


# Step 7
##########################################################################
################
# Decontaminate samples using package "decontam" and method "prevalence",
must have negative control samples
##########################################
```

```
# Uses a chi-square statistic on the 2x2 contingency table of
presence/absence in true samples and in negative controls
# Fisher,Äôs exact test may be used instead of chi-square statistic if
there are too few samples
# Method descriped in :Davis NM, Proctor DM, Holmes SP, Relman DA,
Callahan BJ. Simple statistical identification and removal of contaminant
sequences in marker-gene and metagenomics data. Microbiome. 2018;6: 226.
doi:10.1186/s40168-018-0605-2

# Following tutorital from:
https://benjjneb.github.io/decontam/vignettes/decontam_intro.html

# Inspect library size

df <- as.data.frame(sample_data(ps)) # Put sample_data into a ggplot-
friendly data.frame
df$LibrarySize <- sample_sums(ps)
df <- df[order(df$LibrarySize),]
df$Index <- seq(nrow(df))
ggplot(data=df, aes(x=Index, y=LibrarySize, color=sample_or_control)) +
geom_point()+
  ggsave("library_size.png", path = "")

# Indentify contaminat using the prevalence method

sample_data(ps)$is.neg <- sample_data(ps)$sample_or_control == "control"
contamdf.prev <- isContaminant(ps, method="prevalence", neg= "is.neg")
table(contamdf.prev$contaminant)

head(which(contamdf.prev$contaminant))

contamdf.prev04 <- isContaminant(ps, method="prevalence", neg="is.neg",
threshold=0.35)
table(contamdf.prev04$contaminant)

# Make phyloseq object of presence-absence in negative controls

ps.neg <- prune_samples(sample_data(ps)$sample_or_control == "control",
ps)
ps.neg.presence <- transform_sample_counts(ps.neg, function(abund)
1*(abund>0))

# Make phyloseq object of presence-absence in true positive samples

ps.pos <- prune_samples(sample_data(ps)$sample_or_control == "sample", ps)
ps.pos.presence <- transform_sample_counts(ps.pos, function(abund)
1*(abund>0))

# Make data.frame of prevalence in positive and negative samples

df.pres <- data.frame(prevalence.pos=taxa_sums(ps.pos.presence),
prevalence.neg=taxa_sums(ps.neg.presence),
                      Contaminate=contamdf.prev$contaminant)
```

```
ggplot(data=df.pres,aes(x=prevalence.neg, y=prevalence.pos,
color=Contaminate)) + geom_point()+
  xlab("Prevalence (Negative Controls)") + ylab("Prevalence (Protist
Samples)")+
  ggsave("contam_prev.png", path = "")

# Remove contaminants

ps_noncontam <- prune_taxa(!contamdf.prev04$contaminant, ps)
ps_noncontam


# Remove control samples from phyloseq object

ps_nocontrols <- prune_samples(sample_data(ps_noncontam)$sample_or_control
== "sample", ps_noncontam)
ps_nocontrols


# Step 8
##############################################################################
#######################################
# Perform Taxonomic filtering (i.e what do you expect base on a-priori
knowledge of your samples; Ecological plausibility)
###################################
# Following contamination identification guidelines published in: de
Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS,
et al. Recognizing the reagent microbiome. Nat Microbiol. Springer US;
2018;3: 851,Äì853. doi:10.1038/s41564-018-0202-y
# Create table, number of features for each phyla

table(tax_table(ps_nocontrols)[, "Phylum"], exclude = NULL)

# Remove uncharacterized seqs
ps_nocontrols_b <- subset_taxa(ps_nocontrols, !is.na(Phylum) & !Phylum
%in% c("", "uncharacterized"))
ps_nocontrols_b

# Compute prevalence of each feature, store as data.frame
prevdf = apply(X = otu_table(ps_nocontrols_b),
               MARGIN = ifelse(taxa_are_rows(ps_nocontrols_b), yes = 1, no
= 2),
               FUN = function(x){sum(x > 0)})

# Add taxonomy and total read counts to this data.frame
prevdf = data.frame(Prevalence = prevdf,
                    TotalAbundance = taxa_sums(ps_nocontrols_b),
                    tax_table(ps_nocontrols_b))

#Are there phyla that are comprised of mostly low-prevalence features?
Compute the total and average prevalences of the features in each phylum
plyr::ddply(prevdf, "Phylum",
function(df1){cbind(mean(df1$Prevalence),sum(df1$Prevalence))})
```

```
# Define phyla to filter
filterPhyla = c("Acidobacteria", "Actinobacteria", "Cyanobacteria",
"Euryarchaeota", "Firmicutes", "Fusobacteria", "Gemmatimonadetes",
"Planctomycetes", "Tenericutes", "Chloroflexi", "Chlamydiae")

# Filter entries with Phyla list from above
ps1 = subset_taxa(ps_nocontrols_b, !Phylum %in% filterPhyla)
ps1

# Define Orders to filter
filterOrders = c("Rhizobiales", "Betaproteobacteriales",
"Pseudomonadales", "Sphingomonadales", "Xanthomonadales", "Rickettsiales",
"Rhodobacterales", "Rhodospirillales", "Sphingobacteriales",
"Enterobacteriales", "Salinisphaerales", "Caedibacterales",
"Myxococcales", "Kordiimonadales", "Xanthomonadales", "Flavobacteriales",
"Chitinophagales", "Cytophagales", "Caulobacterales", "Holosporales",
"Oceanospirillales","Oligoflexales", "Paracaedibacterales")

# Filter entries with Order list from above
ps1a = subset_taxa(ps1, !Order %in% filterOrders)
ps1a

# Define Genera to filter
filterGenera = c("Aquabacterium", "Ralstonia", "Sphingomonas",
"Pelomonas", "Escherichia/Shigella", "Bradyrhizobium", "Methylobacterium",
"Stenotrophomonas", "Caulobacter", "Allorhizobium-Neorhizobium-
Pararhizobium-Rhizobium
", "Pelomonas", "Cupriavidus", "Brucella", "Klebsiella", "Curvibacter",
"Vanovorax", "Roseomonas", "Acidibacter", "Vibrionimonas", "Ruegeria")

# Filter entries with Genera list from above
ps1b = subset_taxa(ps1a, !Genus %in% filterGenera)
ps1b

# Step 9
##############################################################################
######################
# Filter based on prevalence in protist samples, i.e remove ASVs in fewer
than X% of samples
# Define prevalence threshold as X% (i.e 5%, 0.05) of total samples

prevalenceThreshold = 0.05263158 * nsamples(ps1b)  # threshold x # samples
= minimum of 3 protist samples
prevalenceThreshold


# Subset to the remaining phyla
prevdf1 = subset(prevdf, Phylum %in% get_taxa_unique(ps1b, "Phylum"))
ggplot(prevdf1, aes(TotalAbundance, Prevalence /
nsamples(ps1b),color=Phylum)) +
  # Include a guess for parameter
  geom_hline(yintercept = 0.1, alpha = 0.5, linetype = 2) +
geom_point(size = 2, alpha = 0.7) +
```

```
  scale_x_log10() +  xlab("Total Abundance") + ylab("Prevalence [Frac.
Samples]") +
  facet_wrap(~Phylum) + theme(legend.position="none") +
  ggsave("tax_prevalence_nonfiltered.png", path = "")

# Execute prevalence filter, using `prune_taxa()` function

keepTaxa = rownames(prevdf1)[(prevdf1$Prevalence >= prevalenceThreshold)]
ps2 = prune_taxa(keepTaxa, ps1b)
ps2


# How many genera would be present after prevalence filtering?

length(get_taxa_unique(ps2, taxonomic.rank = "Genus"))

# Subset to the remaining phyla after prevalence filtering

prevdf2 = subset(prevdf, Phylum %in% get_taxa_unique(ps2, "Phylum"))
ggplot(prevdf2, aes(TotalAbundance, Prevalence /
nsamples(ps2),color=Phylum)) +
  # Include a guess for parameter
  geom_hline(yintercept = 0.1, alpha = 0.5, linetype = 2) +
geom_point(size = 2, alpha = 0.7) +
  scale_x_log10() +  xlab("Total Abundance") + ylab("Prevalence [Frac.
Samples]") +
  facet_wrap(~Phylum) + theme(legend.position="none") +
  ggsave("tax_prevalence_filtered.png", path = "")

# Step 10
##############################################################################
####################

# Transform to relative abundance. Save as new object.

ps2ra = transform_sample_counts(ps2, function(x){(x / sum(x))*100})



#Step 11
##############################################################################
#########
# Remove samples with fewer than X numbers of reads left at this stage of
filtering, remember these are low biomass samples that have been filtered
prior to this step

ps2_low_samples_removed = prune_samples(sample_sums(ps2)>=500, ps2)
ps2_low_samples_removed


#Step 12
##############################################################################
############
```

```
# Filter ASVs that are below a minium relative abundance threshold per
sample

ps_prunedSet =
prune_taxa((taxa_sums(ps2_low_samples_removed)/sum(taxa_sums(ps2_low_sampl
es_removed))) >=1e-5, ps2_low_samples_removed)
ps_prunedSet

#inspect Endomicrobium
endo <- subset_taxa(ps_prunedSet, Genus=="Candidatus_Endomicrobium")
endo

# Extract abundance matrix from the phyloseq object
OTU_endo = as(otu_table(endo), "matrix")
# transpose if necessary
if(taxa_are_rows(endo)){OTU_endo <- t(OTU_endo)}
# Coerce to data.frame
OTU_endo_df = as.data.frame(OTU_endo)
write.table(OTU_endo_df, file = "", sep = " ")

# Transform counts

ps_prunedSet_ra = transform_sample_counts(ps_prunedSet, function(x){(x /
sum(x))*100})


# Step 13
################################################################################
#########
# Ordination of remaining samples and their ASVs
############################################

asv_nmds3 <- ordinate(
  physeq = ps_prunedSet_ra,
  method = "NMDS",
  distance = "bray" # Able to sub in different methods here for distance
ie. "bray" "wunifrac" "unifrac"
)
nmds_plot_pruned = plot_ordination(
  physeq = ps_prunedSet_ra,
  ordination = asv_nmds3,
  label = "cell_id",
  color = "Host_genus",
  shape = "termite_colony",
  title = "Ordination of Bacterial Communities of Hindgut Protists (Bray-
Curtis)"
)+
  stat_ellipse(type = "t", linetype = 5) +                          #type =
"t" = multivariate t distribution, can also be "norm"
  labs(col = "Protist Genera", shape = "Termite Colonies")+
  ggsave("protists_nmds_bray_pruned.png", path = "" )
print(nmds_plot_pruned)
```

```
# Step 14
#########################################################################
############################
# Make heatmap for samples and thier ASVs, separate by Protist Genus

# Make heatmap of pruned set
                                  #ps_prunedSet_B or ps_prunedSet_ra
heatmap_prunedSet = plot_heatmap(ps_prunedSet_ra, "NMDS", "bray",
"cell_id", "Phylum", na.value = "grey90", low="#000033", high="#66CCFF")+
  geom_tile(colour="white",size=0.40)+
  #set a base size for all fonts
  theme_grey(base_size=8)+
  #theme options
  theme(
    #bold font for both axis text
    axis.text=element_text(face="bold"),
    #set thickness of axis ticks
    axis.ticks=element_line(size=0.4),
    #remove plot background
    plot.background=element_blank(),
    #remove plot border
    panel.border=element_blank())+
  facet_grid(~Host_genus, scale="free")+
  theme(strip.text.x = element_text(size = 12))+
theme(axis.text.x = element_text(size = 7, angle = 45))+
theme(legend.position = "bottom")
heatmap_prunedSet$scales$scales[[1]]$name <- "Protist Cells"
heatmap_prunedSet$scales$scales[[2]]$name <- "Bacterial ASVs (Phylum)"
print(heatmap_prunedSet)+
  ggsave("heatmap_prunedSet_ra.png", path = "" )


# Step 15
#########################################################################
###################
# Estimate and Plot Richness/ Alpha diversity #######

ps_prunedSet_alpha_est <- estimate_richness(ps_prunedSet, measures =
c("Shannon", "Chao1", "Observed"))
write.csv(ps_prunedSet_alpha_est, file ="" )

alpha_p <- plot_richness(ps_prunedSet, x="Host_genus",
measures=c("Observed", "Shannon", "Chao1"), color="termite_colony")+
  labs(col = "Termite Colonies", x = "Protist Genera")+
  ggsave("protist_richness_plots.png", path = "" )
(alpha_box <- alpha_p + geom_boxplot(data=alpha_p$data, aes(x =
Host_genus, y=value, color=NULL), alpha=0.1))+
  ggsave("protist_richness_boxplots.png", path = "" )


#Step 16
#########################################################################
###################
# Arrange all final graphs (Heatmap, NMDS, Abundance) together
```

```
# Make figure

g_heatmap_plots_prunedSet <-arrangeGrob(heatmap_prunedSet, alpha_box,
nmds_plot_pruned,
                                        ncol = 3, nrow = 2,
                                        layout_matrix = rbind(c(1,2),
c(1,3), c(1,4)))
ggsave("heatmap_plots_prunedSetB.png", g_heatmap_plots_prunedSet, path =
"")

# Step 19
########################################################################
#######################
# Run Adonis statistic on samples (Is there a difference between the
protist Genera?)
# Extract Abundance matrix from phyloseq object

ASV_table_pruned = OTU1 = as(otu_table(ps_prunedSet_ra), "matrix")

# Transpose ASV table if needed

if(taxa_are_rows(ps_prunedSet_ra)){ASV_table_pruned <-
t(ASV_table_pruned)}

# Coerce to data.fram

ASV_table_pruned_df = as.data.frame(ASV_table_pruned)
write.table(ASV_table_pruned_df, "", sep="\t", quote=F)


# Run adonis on ASV dataset
ASV_dist = phyloseq::distance(ps_prunedSet_ra, "bray")
ps_prunedSet_NMDS = ordinate(ps_prunedSet_ra, "NMDS", ASV_dist)
adonis(ASV_dist ~ Host_genus, as(sample_data(ps_prunedSet_ra),
"data.frame"))

adonis(ASV_dist ~ termite_colony, as(sample_data(ps_prunedSet_ra),
"data.frame"))

ps_trichonympha = subset_samples(ps_prunedSet_ra, Host_genus==
"Trichonympha")
ps_trichonympha

ASV_dist = phyloseq::distance(ps_trichonympha, "bray")
ps_prunedSet_NMDS = ordinate(ps_trichonympha, "NMDS", ASV_dist)
adonis(ASV_dist ~ termite_colony, as(sample_data(ps_trichonympha),
"data.frame"))

ps_pyrsonympha = subset_samples(ps_prunedSet_ra, Host_genus==
"Pyrsonympha")
ps_pyrsonympha

ASV_dist = phyloseq::distance(ps_pyrsonympha, "bray")
ps_prunedSet_NMDS = ordinate(ps_pyrsonympha, "NMDS", ASV_dist)
```

```
adonis(ASV_dist ~ termite_colony, as(sample_data(ps_pyrsonympha),
"data.frame"))

ps_dinenympha = subset_samples(ps_prunedSet_ra, Host_genus== "Dinenympha")
ps_dinenympha

ASV_dist = phyloseq::distance(ps_dinenympha, "bray")
ps_prunedSet_NMDS = ordinate(ps_dinenympha, "NMDS", ASV_dist)
adonis(ASV_dist ~ termite_colony, as(sample_data(ps_dinenympha),
"data.frame"))


# Make bar charts

bar_plot = plot_bar(ps_prunedSet_ra, "cell_id", fill="Phylum")+
           facet_grid(~Host_genus, scale="free")
print(bar_plot)
ggsave("barplot_prunedSet.png", bar_plot, path = "")


#inspect Treponema ASVs
spiro_asvs <- subset_taxa(ps_prunedSet, Phylum=="Spirochaetes")
spiro_asvs

# Extract abundance matrix from the phyloseq object
OTU_spiro = as(otu_table(spiro_asvs), "matrix")
# transpose if necessary
if(taxa_are_rows(spiro_asvs)){OTU_spiro <- t(OTU_spiro)}
# Coerce to data.frame
OTU_spiro_df = as.data.frame(OTU_spiro)
write.table(OTU_spiro_df, file = "", sep = " ")


#inspect Candidatus_Symbiothrix ASVs
bacteroid_asvs <- subset_taxa(ps_prunedSet, Phylum=="Bacteroidetes")
bacteroid_asvs

# Extract abundance matrix from the phyloseq object
OTU_bacteroid = as(otu_table(bacteroid_asvs), "matrix")
# transpose if necessary
if(taxa_are_rows(bacteroid_asvs)){OTU_bacteroid <- t(OTU_bacteroid)}
# Coerce to data.frame
OTU_bacteroid_df = as.data.frame(OTU_bacteroid)
write.table(OTU_bacteroid_df, file = "", sep = " ")

#inspect Desulfovibrio ASVs
proteo_asvs <- subset_taxa(ps_prunedSet, Phylum=="Proteobacteria")
proteo_asvs

# Extract abundance matrix from the phyloseq object
OTU_proteo = as(otu_table(proteo_asvs), "matrix")
# transpose if necessary
if(taxa_are_rows(proteo_asvs)){OTU_proteo <- t(OTU_proteo)}
# Coerce to data.frame
```

```
OTU_proteo_df = as.data.frame(OTU_proteo)
write.table(OTU_proteo_df, file = "", sep = " ")

#inspect Verrucomicrobia ASVs
Verrucomicrobia_asvs <- subset_taxa(ps_prunedSet,
Phylum=="Verrucomicrobia")
Verrucomicrobia_asvs

# Extract abundance matrix from the phyloseq object
OTU_Verrucomicrobia = as(otu_table(Verrucomicrobia_asvs), "matrix")
# transpose if necessary
if(taxa_are_rows(Verrucomicrobia_asvs)){OTU_Verrucomicrobia <-
t(OTU_Verrucomicrobia)}
# Coerce to data.frame
OTU_Verrucomicrobia_df = as.data.frame(OTU_Verrucomicrobia)
write.table(OTU_Verrucomicrobia_df, file = "", sep = " ")

#inspect Margulisbacteria ASVs
Margulisbacteria_asvs <- subset_taxa(ps_prunedSet,
Phylum=="Margulisbacteria")
Margulisbacteria_asvs

# Extract abundance matrix from the phyloseq object
OTU_Margulisbacteria = as(otu_table(Margulisbacteria_asvs), "matrix")
# transpose if necessary
if(taxa_are_rows(Margulisbacteria_asvs)){OTU_Margulisbacteria <-
t(OTU_Margulisbacteria)}
# Coerce to data.frame
OTU_Margulisbacteria_df = as.data.frame(OTU_Margulisbacteria)
write.table(OTU_Margulisbacteria_df, file = "", sep = " ")

# Write taxonomy file
tax_prundedset <- tax_table(ps_prunedSet)

write.csv(tax_prundedset, file = "", sep = " " )

# inspect Synergistetes ASVs
Synergistetes_asvs <- subset_taxa(ps_prunedSet, Phylum=="Synergistetes")
Synergistetes_asvs

# Extract abundance matrix from the phyloseq object
OTU_Synergistetes = as(otu_table(Synergistetes_asvs), "matrix")
# transpose if necessary
if(taxa_are_rows(Synergistetes_asvs)){OTU_Synergistetes <-
t(OTU_Synergistetes)}
# Coerce to data.frame
OTU_Synergistetes_df = as.data.frame(OTU_Synergistetes)
write.table(OTU_Synergistetes_df, file = "", sep = " ")

#inspect Endomicrobium
endo2 <- subset_taxa(ps_prunedSet, Genus=="Candidatus_Endomicrobium")
endo2

# Extract abundance matrix from the phyloseq object
```

```
OTU_endo2 = as(otu_table(endo2), "matrix")
# transpose if necessary
if(taxa_are_rows(endo2)){OTU_endo2 <- t(OTU_endo2)}
# Coerce to data.frame
OTU_endo2_df = as.data.frame(OTU_endo2)
write.table(OTU_endo2_df, file = "", sep = " ")


# Write taxonomy file
tax_prundedset <- tax_table(ps_prunedSet)

write.csv(tax_prundedset, file = "", sep = " " )
```

Chapter Three

Endosymbionts of Protist Hosts Use Gene Flow to Acquire Niche Specific Traits in the Nested Symbiosis of Termites[+]

Contribution from other researchers

Dr. Jacquelynn Benjamino contributed to this chapter. She performed whole genome and transcriptome amplification as well as prepared sequencing libraries. Dr. Benjamino also performed transcriptome analysis and contributed to the discovery of the differences in the carbon metabolisms of the *Endomicrobium* species.

[+] In preparations for submission, Stephens, Benjamino, Graf, and Gage.

Abstract

Different protist species which colonize the hindguts of wood-feeding *Reticulitermes* termites are associated with endosymbiotic bacteria belonging to the genus *Endomicrobium*. In this study, we focused on the endosymbionts of three protist species from *Reticulitermes flavipes*, which included *Pyrsonympha vertens*, *Trichonympha agilis*, and *Dinenympha* species II. Since these protist hosts represented members of difference taxa which colonize different niches within the hindguts of their termite hosts, we investigated if these differences translated to differential gene content and expression in their endosymbionts. Following assembly and comparative genome analyses, we discovered that these endosymbionts differed with respect to several niche specific traits such as, aerotolerance and carbon metabolism. Our analyses supported that genes related to carbon metabolism were acquired by horizontal gene transfer (HGT) from donor taxa which are present in termite's hindgut community. In addition, our analyses supported that these endosymbionts have retained and expressed several genes related to natural transformation (competence) and recombination. Taken together, the presence of genes acquired by HGT and a putative competence pathway supported that these endosymbionts are not cut-off from gene flow and that competence may be a mechanism by which members of the *Endomicrobium* can acquire new traits.

Introduction

Among the wood-feeding lower termites, symbiotic protists which reside in the hindgut are often colonized by endosymbionts (Hongoh et al., 2008b, 2008a; Ohkuma, 2008; Strassert et al., 2016). In *Reticulitermes* spp. both Oxymonadida (order) and Parabasilia (phylum or class)

protists associate with endosymbiotic bacteria belonging to the genus *Endomicrobium* (phylum Elusimicrobia) (Ohkuma, 2008; Ohkuma et al., 2007; Stingl et al., 2005). Members of *Endomicrobium* have been shown to comprise a significant portion of the core bacterial community in wood-feeding termites such as *Reticulitermes flavipes* (Benjamino and Graf, 2016; Boucias et al., 2013). The endosymbiotic lineages began their associations with hindgut protists approximately 70 - 40 million years ago (mya) (Ikeda-Ohtsubo and Brune, 2009) and arose from free-living relatives during multiple independent acquisition events (Mikaelyan et al., 2017). Since their intimate partnership with protist hosts was established, the serial vertical passage from one protist cell to its progeny have enabled the two partners to co-speciate, as inferred by congruent ribosomal RNA (rRNA) phylogenies (Ikeda-Ohtsubo et al., 2007; Ikeda-Ohtsubo and Brune, 2009; Zheng et al., 2015).

Several *Endomicrobium* lineages occupy different niches within the hindguts of termites. In addition to colonizing the cytoplasm of certain hindgut protist species, *Endomicrobium* spp. have been observed as ectosymbionts of protists (Izawa et al., 2017), as well as free-living (not protist-associated) (Ikeda-Ohtsubo et al., 2010; Mikaelyan et al., 2017; Zheng et al., 2016). Because of their distribution across these different niches within termite hosts, they may provide an important opportunity for studying bacterial genome evolution across different symbiotic states (free-living, endosymbiotic, and ectosymbiotic).

A previous study performed comparative genome analyses of both a putatively free-living *Endomicrobium* (PFLE), *Endomicrobium provatium* Rsa215 (Zheng et al., 2016) and an endosymbiont '*Candidatus* Endomicrobium trichonymphae' Rs-D17, of a *Trichonympha* protist species (Hongoh et al., 2008a) to determine differences between these two lineages (Zheng et al., 2017). Their findings suggested that the transition from PFLE to an intracellular lifestyle

involved genome reduction, similar to that of endosymbionts of sap-feeding insects, but also the incorporation of genes, possibly from termite gut co-inhabitants, by horizontal gene transfer (HGT) (Zheng et al., 2017). The genome of '*Ca*. E. trichonymphae' Rs-D17 appeared to have acquired several adaptions by HGT which support the nutrient provisioning of its protist host. These include sugar and amino acid transporters, and other genes involved in amino acids biosynthesis (Zheng et al., 2017). These findings suggested that unlike the endosymbionts of sap-feeding insects, *Endomicrobium* may not be cut-off from gene flow (Zheng et al., 2017).

In this study, we expand upon that, and similar studies findings by presenting and comparing near-complete draft genomes (along with transcriptome data) of three *Endomicrobium* species, assembled from single protist cell metagenomes of three protists species that co-inhabit the hindgut of *R. flavipes*. Collectively, these endosymbionts represented populations exposed to different environments as one protist host (*Pyrsonympha vertens*) lives attached to the oxic gut wall (Brune et al., 1995; Yang et al., 2005), while the other two protist species are motile and found in the anoxic hindgut lumen (*Trichonympha agilis* and *Dinenympha* species II). In addition, *P. vertens* and *D.* species II are both Oxymonads while *T. agilis* is a Parabasalid. This allowed us to investigate differences between their endosymbionts in the context of the relatedness between protist hosts.

Our findings supported that these different *Endomicrobium* species have differences in gene content related to aerotolerance and oxidative stress response, as well as carbon usage and metabolism. Interestingly, as seen previously in '*Ca*. E. trichonymphae' Rs-D17, these *Endomicrobium* species have also acquired genes by HGT from putative donor taxa which have termite-associated members. In addition, we describe data suggesting that *Endomicrobium* have retained a competence pathway which may have allowed them to import and incorporate

exogenous DNA into their genomes. Genes involved in this pathway were both conserved and expressed in these endosymbiont of hindgut protists.

Methods

Termite Collection and Species Identification

Termites were collected from Mansfield Connecticut and their identity was verified as previously described (Stephens and Gage, 2018) by amplifying and sequencing the mitochondrial cytochrome oxidase II gene. Termites were maintained in the lab with sterile moistened sand, that was initially sterilized, and spruce wood.

Single Protist Cell Isolation

Worker termites were brought into an anaerobic chamber and their hindguts were dissected with sterile forceps. Hindguts were ruptured in ice-cold Trager's Solution U (TU) (Trager, 1934) and washed three times at 3,000 rpms for 90 seconds. These samples were then diluted in ice-cold TU and spotted onto glass slides which were treated with RNase AWAY™ (Thermo Fisher Scientific) and UV light. Single protist cells were isolated by micromanipulation using an Eppendorf CellTram® Vario equipped with a sterile glass capillary tube which had been treated with UV light, 70% ethanol, and RNase AWAY™. Individual protist cells were washed three times in ice-cold TU by micromanipulation, placed in 10 µl of ice-cold molecular grade phosphate buffered saline (PBS), flash frozen on dry ice, and immediately stored at -80 degrees Celsius.

Whole Genome and Transcriptome Amplification

The metagenome (DNA) and metatranscriptome (cDNA) from the protist cell samples were amplified 12-24 hours after isolation. Cell lysis and amplification was performed using the Repli-g WGA/WTA kit (Qiagen). Cells were lysed using a Qiagen lysis buffer and incubation step, immediately followed by incubation on ice. Two samples from the same lysed cell were taken and used in simultaneous whole genome amplification and whole transcriptome amplification. These were carried out per the kits' standard protocol with exception of the addition of random hexamer primers which were used to amplify DNA and cDNA.

Library Preparation and Sequencing

Both DNA and cDNA were sheared using a Covaris M220 ultra-sonicator™ according to standard protocol. WGA samples were sheared to a 550 bp insert size using 200 ng of DNA and WTA samples were sheared to a 350 bp insert size using 100 ng of cDNA. Sequencing libraries were prepared using the TruSeq Nano DNA Library Prep kit from Illumina® according to standard protocol. Each sample was prepared with a forward and reverse barcode for the ability to add all samples on the same sequencing run. The samples were sequenced using an Illumina® NextSeq 1x150 Mid Output run and two NextSeq 1x150 High Output runs.

Genomic Read Processing and Assembly

Reads were preprocessed before assembly with a BBmap (Bushnell, 2014) workflow. Reads were filtered for contaminating sequences by mapping reads to reference genomes of potential contamination sources such as human DNA, human associated microbiota, and organisms commonly used in our research laboratories. A list of references genomes used for contamination filtering is provided as a Supplementary Table 1. Using the "bbduck.sh script",

adaptor sequences were trimmed from reads, and last base pair of 151 bp reads were removed using the "ftm" command. Reads were trimmed at both ends using the quality score of Q15 as a cutoff. Homopolymers were removed by setting an entropy cutoff of 0.2, a max G+C cutoff of 90%, and by removing reads which possessed stretches of G's equal to or greater than 25 bases long. In addition, reads which were below a minimum average quality of Q15 and/or 50 bases long were removed. Genomic reads were then normalized to a minimum coverage of 2X and a maximum coverage of 50X and then deduplicated using BBnorm. Genomic reads were then assembled using the A5 assembly pipeline (Coil et al., 2015) on the KBase web server (Arkin et al., 2018).

Genomic Binning, Draft Genome Assessment, and Annotation

Metagenomic assemblies from single protist host cells and their bacterial symbionts were binned using either 4-mer or 6-mer frequencies with VizBin (Laczny et al., 2015) using scaffolds of at least 1Kb in size. Genomic bins of interest (low GC content) were selected. Each scaffold from these bins were used in a BLAST+ search (blastn) (Camacho et al., 2009) against previously sequenced Elusimicrobia genomes (references used are given in Supplementary Table 2). Scaffolds which had a positive hit to other Elusimicrobia (at least 70% identity over a 1kb alignment) were retained in the draft genomes and scaffolds which did not have a significant hit to other Elusimicrobia genomes were used in a second BLAST+ search against the non-redundant (NR) database. Scaffolds which had positive hits to other Elusimicrobia sequences such as, sequences derived from metagenome assemblies and/or other experimental sequences, were retained in the draft genomes. Draft genomes were then iteratively polished with the program Pilon (Walker et al., 2014). These draft genomes were then assessed for contamination

and completeness using CheckM (Parks et al., 2015) which uses lineage specific marker genes to perform the analyses. The resulting near complete draft genomes were then annotated on the RAST Server (Aziz et al., 2008) using a customized RASTtk (Brettin et al., 2015) workflow with options selected to call insertion sequences and prophages. Metabolic pathways pertaining to carbon metabolism, amino acid biosynthesis, vitamin biosynthesis, and peptidoglycan biosynthesis were reconstructed using described pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000).

Analysis of Ribosomal Gene Phylogeny and Average Nucleotide Identities

Ribosomal 16S genes from each of the *Endomicrobium* spp. draft genomes were trimmed and aligned to references using MUSCLE (Edgar, 2004), evolutionary models were tested and a Maximum likelihood (ML) phylogenetic tree was made using IQ-TREE (Nguyen et al., 2015). JSpeciesWS (Richter et al., 2016) was used for determining the average nucleotide identities based on BLAST+ search (ANIb) between the *Endomicrobium* spp. draft genomes and the genome of '*Ca*. Endomicrobium trichonymphae' Rs-D17, which is a close relative (Hongoh et al., 2008a).

Assembled 18S rRNA genes were retrieved from metagenome assemblies by performing a BLAST+ search using previously published 18S rRNA reference sequences for each protist species as queries (Stephens and Gage, 2018). When possible, protist 18S rRNA genes were amplified using leftover DNA from WGA samples using universal primers 18SFU; 5'-ATGCTTGTCTCAAAGGRYTAAGCCATGC-3' and 18SRU; 5'-CWGGTTCACCWACGGAAACCTTGTTACG-3' (Tikhonenkov et al., 2016) as previously described (Stephens and Gage, 2018) and sequenced by sanger sequencing. This confirmation

PCR was done using samples TA21, TA26, and DS12. Assembled 18S rRNA genes were aligned to references using MUSCLE and a Maximum likelihood (ML) phylogenetic tree was generated using IQ-TREE with model testing.

Detection of Horizontally Acquired Genes

Genes acquired by horizontal gene transfer were detected using phylogenetic methods. Initially, protein sequences of genes of interest that were not shared across the different *Endomicrobium* spp. were aligned to references that spanned different bacterial taxa using MUSCLE and phylogenetic trees were generated using IQ-TREE with model testing. Gene trees were then compared to the 16S rRNA gene tree phylogeny (Supplemental Figure 4) to determine incongruence between their evolutionary histories.

Analysis of Genes Involved in Competence and Recombination

Genes known to be involved in DNA uptake, competence, and recombination were obtained from each *Endomicrobium* spp. draft genome based on their RAST annotations and homology to reference sequences. The distribution of these genes was then compared across the different draft genomes and references which included free-living relatives and other endosymbionts. To asses if these genes were complete and retained their putative functions, homologs of each gene were obtained from genomes of bacteria belonging to the Phylum Elusimicrobia, aligned with MUSCLE, and phylogenetic trees were generated using IQ-TREE (Nguyen et al., 2015) with model testing, and support values generated using the "–abayes" and "–bb 1000" commands. The resulting phylogenetic trees were used along with the MUSCLE alignments to perform a dN/dS analysis using the program Codeml which is a part of the PAML

and PAMLX packages (Xu and Yang, 2013; Yang, 1997). The presence other genes such those involved in pilus formation and secretion systems was investigated using the program Macromolecular Systems Finder (MacSyFinder) (Abby et al., 2014).

Mapping Transcriptome Reads to Draft Genomes

RNA-seq (metatranscriptome) reads were quality trimmed and filtered as described above and rRNA reads were removed by mapping to references (references given as supplementary table 1). Transcriptome reads were error corrected in Geneious R11 (Kearse et al., 2012) using BBNorm with default settings. Reads were then mapped to their respective *Endomicrobium* spp. draft genome in Geneious R11 using Bowtie 2 (Langmead and Salzberg, 2012) with alignment type set to "End to End" and using the "Medium Sensitivity" preset. Expression levels were then calculated in Geneious R11, excluding ambiguously mapped reads.

Verification of *comEC* Expression by Reverse Transcriptase PCR (PT-PCR)

Primers were designed to amplify *comEC* from '*Ca*. Endomicrobium agilae' in Geneious R11 using Primer3 (Untergasser et al., 2012) (Primers: endo_comec_F: 5'-ATTTGCCTGTGTTTGAGAGT-3' and endo_comec_R: 5'-CCTGTTCCTGTGCTTTCAG -3').Twenty termites were used to prepare RNA and cDNA samples for reverse transcriptase PCR (RT-PCR) reactions. Termite hindguts were dissected and ruptured in TU on ice in an anaerobic chamber. Hindgut contents were washed with ice-cold TU three times at 3,000 rpms for 90 seconds and then lysed in 1mL of TRIzol™ Reagent (Thermo Fisher Scientific), and RNA was isolated per the manufacturer's protocol. RNA was then treated with TURBO™ DNase (Thermo Fisher Scientific) following the manufacturer's protocol for 50µl reactions using 20µl (~5.6mg)

of total RNA. TURBO™ DNase-treated RNA was then used as template for cDNA synthesis using SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) following the manufacturer's protocol for 20µl, first strand synthesis reactions using random hexamer primers. The resulting cDNA was then treated with *E. coli* RNase H for 20 minutes at 37 degrees Celsius.

RT-PCR reactions were performed using the *Endomicrobium comEC* primers with RNase-treated cDNA serving as template and no-RT control template consisting of DNase-treated RNA. RT-RCR was performed using Phusion Polymerase (Thermo Fisher Scientific) with HF buffer and DMSO. Cycling conditions were as follows: initial denaturation at 94 degrees C for 3 minutes, followed by 35 cycles of 94 degrees C for 45 seconds, annealing at 59 degrees C for 30 seconds, and extension at 72 degrees C for 45 seconds. Final extension was done 72 degrees C for 10 minutes. Hindgut DNA (washed protists cell fractions from five hindguts in molecular grade Tris EDTA buffer) was used as a positive PCR control. RT-PCR products were visualized using a 1% agarose gel with ethidium bromide, then purified from gel slices using the Monarch DNA gel purification kit (New England Biolabs), and sequenced by Sanger sequencing.

Results

Phylogeny of Protist Hosts

Protist 18S rRNA genes were retrieved from metagenome assemblies and confirmed independently by PCR and Sanger sequencing. A Maximum likelihood (ML) phylogenetic tree was made, and the resulting phylogenies supported that the species identification of the protist cells used in this study were *Trichonympha agilis* (cells TA21 and TA26), *Pyrsonympha vertens* (cells PV1 and PV7), and *Dinenympha* species II (cell DS12) (Figure 1).

Endomicrobium Genome Statistics and Speciation

Five near-complete *Endomicrobium* genomes were obtained from five single protist cell metagenomic assemblies. Each of these five genomes ranged from 1.12 - 1.37 mb in size, 35.3 – 36.6 % G+C, and 93.3 - 96.6% completeness (Figure 2A). To determine if these genomes were from the same of different *Endomicrobium* species, we calculated pairwise Average Nucleotide Identities (ANI) with an ANI score of 95% or greater as a biomarker for a species-level cutoff (Jain et al., 2018). From *T. agilis* samples, we assembled two draft genomes which had and ANI score of greater than 95% to one another but less than 95% to '*Ca*. E. trichonymphae' Rs-D17 (Figure 2B). Based on this analysis, we designated a new Candidatus species name for these organisms and refer to them as '*Candidatus* Endomicrobium agilae' TA21 and TA26. We also assembled two *Endomicrobium* genomes from *P. vertens* samples which had an ANI score of greater than 95% to one another but less than 95% to other *Endomicrobium* genomes (Figure 2B). The 16S rRNA genes from these genomes were greater than 98.6% identical to a previously described species, '*Candidatus* Endomicrobium pyrsonymphae' (Stingl et al., 2005), which is the Candidatus species designation that we use for genomes PV1 and. One additional *Endomicrobium* genome was assembled from *Dinenympha* species II cell DS12. This genome did not share an ANI score greater than 95% to other *Endomicrobium* genomes and was thus given a new Candidatus species designation hereinto referred as '*Candidatus* Endomicrobium dinenymphae' DS12 (Figure 2B).

Collectively these *Endomicrobium* genomes contained between 1005 – 1230 orthologous gene clusters (OGCs), of which 717 were shared with one another (Figure 2C). Additionally, 409 OCGs were unique to '*Ca*. E. agilae' TA21 and TA26 and another 183 OGCs were unique to '*Ca*. E. pyrsonymphae' PV1 and PV7 (Figure 2C). Although the genome of 'Ca. E.

dinenymphae' DS12 only had 24 unique OGCs, it shared 153 with '*Ca*. E. pyrsonymphae' PV1 and PV7 (Figure 2C) which may reflect similar selective pressure for gene retention between endosymbionts of Oxymonad hosts, or their more recent shared history, compared to the *Endomicrobium* associated with Parabasalid hosts.

Biosynthesis of Amino Acids, Vitamins, and Peptidoglycan

In general, each of the five *Endomicrobium* genomes obtained in this study contained similar genes involved in the biosynthesis of amino acids (Supplemental Figure 1A), vitamins (Supplemental Figure 1B), and peptidoglycan (Supplemental Figure 2). Each of the *Endomicrobium* genomes possessed completed pathways for arginine, valine, leucine, isoleucine, and tryptophan biosynthesis (Supplemental Figure 1A). Interestingly, the *Endomicrobium* symbionts (PV1, PV7, and DS12) of Oxymonad protists lacked at least one gene in each of the biosynthesis pathways for histidine, phenylalanine, and threonine which were all present in the genomes of '*Ca.* E. agilae' TA21 and TA26 (Supplemental Figure 1A). Each genome lacked several of genes related to lysine biosynthesis (Supplemental Figure 1A), however it has been hypothesized that *Endomicrobium* may use an alternative pathway for the synthesis of this amino acid (Hongoh et al., 2008a). None of these five genomes encoded genes related to the biosynthesis of cysteine or methionine, however PV1, PV7, and DS12 encoded a Methionine transporter (MetT) (Supplemental Figure 1A).

Regarding the other amino acids, each of the *Endomicrobium* genomes contained completed pathways for the biosynthesis of aspartic acid, tyrosine, glutamine, glutamate, and glycine (Supplemental Figure 1A). In addition, each genome encoded a serine transporter but only PV1, PV7, and DS12 encoded a glutamate transporter (Supplemental Figure 1A). The

genomes for '*Ca.* E. agilae' TA21 and TA26 were missing at least one gene in the Proline biosynthesis pathway, which was completed in the other *Endomicrobium* genomes (PV1, PV7, and DS12) (Supplemental Figure 1A). None of these *Endomicrobium* genomes encoded genes involved in the biosynthesis of asparagine or alanine (Supplemental Figure 1A).

The content of genes involved in the biosynthesis of vitamins and co-factors were also similar in these *Endomicrobium* genomes, with most being incomplete (Supplemental Figure 1B). For example, each genome encoded a near-complete Biotin biosynthesis pathway, missing just a single gene (*bioW*) needed to convert pimelate to Biotin (Supplemental Figure 1B). Several genes in the thiamine biosynthesis pathway (pyruvate to thiamine) were also missing in each of these genomes (Supplemental Figure 1B). Each *Endomicrobium* genome encoded a Vitamin B12 transporter (*btuC*) and the necessary enzyme needed to convert Vitamin B12 to FMN and FAD (Supplemental Figure 1B). With the exception of DS12, all genomes possessed completed pathways to synthesize NAD+ and NADP+ and none possessed the completed pathways for Lipoic Acid synthesis (Supplemental Figure 1B). Regarding peptidoglycan synthesis, each *Endomicrobium* genome encoded the same incomplete pathway with genes *dgkA* (undecaprenol kinase) and *vanY* (D-alanyl-D-alanine carboxypeptidase) missing (Supplemental Figure 2).

Oxidative Stress Response and Aerotolerance

Since the protist hosts of these endosymbionts occupy two distinct niches within the termite hindgut environment (attached to the oxic gut wall (*P. vertens*) versus being motile within the anoxic lumen) (Figure 3A), we investigated differences related to genes involved aerotolerance and oxidative stress responses in the *Endomicrobium* genomes. Each genome encoded rubredoxin, rubredoxin-oxygen oxidoreductase, and rubrerythrin (Figure 3B) which

have been shown to confer oxidation stress protection in the anaerobic bacterium *Desulfovibrio vulgaris* (Lumppio et al., 2001).

These genomes also contained genes of the Bacteroidetes aerotolerance operon (*Batl*), which is an operon that consists of total of five genes (*batA-E*), which collectively encode a complex that confers aerotolerance (Tang et al., 1999). Interestingly, only '*Ca*. E. pyrsonymphae' PV1 and PV7, whose hosts are attached to the oxic gut wall of their termite host, contained all five genes of this aerotolerance operon (Figure 3B). In contrast, the other *Endomicrobium* genomes were each missing a single gene of this operon (Figure 3B). The competed *Batl* operon was encoded in the genome of '*Ca*. E. trichonymphae' Rs-D17, which is the closest sequenced relative to '*Ca*. E. agilae' TA21 and TA27. We mapped metagenomic reads from samples TA21 and TA26 to the reference '*Ca*. E. trichonymphae' Rs-D17 genome and then inspected the *Batl* operon coverage to determine if the *batB* from '*Ca*. E. agilae' TA21 and TA27 could have been missed as an artifact of assembly or compositional binning. The resulting mapping data supported that the *batB* gene was absent in samples TA21 and TA26 as no reads mapped to that region in the reference '*Ca*. E. trichonymphae' Rs-D17 genome. The distribution of genes in the *Batl* operon in the *Endomicrobium* genomes presented in this study are consistent with their protist hosts' exposure to the oxic or anoxic regions of the termite hindgut (Figure 3).

Differences in Carbon Metabolism

Some interesting differences between these *Endomicrobium* genomes pertain to their carbon metabolism. Each of these three *Endomicrobium* species had relatively simple pathways for importing and using different wood-derived carbon sources. For example, '*Ca.* E. agilae'

TA21 and TA26 encoded all the genes necessary to import and use both glucuronate and glucose-6-phosphate (Figure 4A & 4B), which are the same carbon sources that can be used by '*Ca*. E. trichonymphae' Rs-D17 (Hongoh et al., 2008a). Interestingly, genome analyses suggest that these two carbon sources cannot be used by the other *Endomicrobium* species which instead use either arabinose ('*Ca*. E. pyrsonymphae' PV1 and PV7) or xylose ('*Ca*. E. dinenymphae' DS12) as their sole carbon source (Figures 5A & 5B; Figures 6A & 6B, respectively). Transcriptome data supported that each of the genes involved in these carbon usage pathways were expressed in the respective *Endomicrobium* while residing in their protist hosts (Figures 4C, 5C, & 6C). Metabolites from these carbon sources are fed into both the non-oxidative pentose phosphate pathway and glycolysis (Figures 4B, 5B, 6B, & Supplemental Figure 3).

Other differences in carbon metabolism included their fermentation end products (Supplemental Figure 3). Following glycolysis, pyruvate can be fermented to acetate by all three of the *Endomicrobium* species, however ethanol can also be produced by '*Ca.* E. agilae' and '*Ca*. E. dinenymphae', but not by '*Ca*. E. pyrsonymphae' (Figures 4B, 5B, & 6B). In addition, lactate can be produced by both '*Ca*. E. dinenymphae', and '*Ca*. E. pyrsonymphae', but not '*Ca.* E. agilae' (Figures 4B, 5B, & 6B).

Since previous studies identified genes acquired by horizontal gene transfer (HGT) in other *Endomicrobium* species (Zheng et al., 2017), we tested whether HGT could, at least in part, explain the differences seen in carbon metabolism across the genomes presented in this study. For each of carbon usage pathways for glucuronate, arabinose, and xylose, phylogenetic trees were made from amino acids alignments of the transporters and isomerases (Figures 4D, 5D, & 6D) and the phylogenies were compared to the *Endomicrobium* 16S rRNA gene phylogeny (Supplemental Figure 4) to determine if they were congruent or not. In each case, these

phylogenies were incongruent, supporting that these gens were acquired by HGT from putative donor taxa which are present in the gut of termites (Figures 4D, 5D, & 6D), similar to previous reports of HGT in '*Ca*. E. trichonymphae' Rs-D17 (Zheng et al., 2017). These donor taxa included Bacteroidetes, Actinobacteria, and Firmicutes (Figures 4D, 5D, & 6D), which are all part of the hindgut community of *R. flavipes* (Benjamino and Graf, 2016). These data supported previous conclusions by others that *Endomicrobium* are not cut off from gene flow (Zheng et al., 2017). This is in contrast to the endosymbionts of sap-feeding insects, which are traditionally thought to experience little to no gene flow, however, recent analyses suggested that HGT may occur more frequently than previously thought (López-Madrigal and Gil, 2017).

Natural Transformation (Competence) as a Putative Mechanism for Acquiring Genes

All currently sequenced genomes of endosymbiotic *Endomicrobium* lineages have been shown to have acquired genes by HGT, suggesting that their gene content could reveal insights into a putative conserved mechanism by which these genes were acquired. Interestingly, compared to other endosymbiotic lineages, the *Endomicrobium* genomes were relatively enriched in genes related to the uptake and recombination of exogenous DNA (Supplemental Figure 5). Of special interest are the *Endomicrobium* genes *comEC*, *comF*, *ssB*, *drpA*, and *recA* which are all involved in natural transformation (competence) in bacteria such as *Vibrio cholerae* (Seitz and Blokesch, 2013).

These genes were further investigated by determining if selection was acting on them by measuring their dN/dS ratios. In addition, we measured transcription of these genes as well. The dN/dS analyses of these genes supported that selection was acting to maintain the amino acid sequences of their corresponding gene products (dN/dS < 1.0, with the exception of *ssb* from

TA21) (Figure 7A) and transcriptome analysis supported that they were expressed (Figure 7B). The gene *comEC*, which encodes a transporter that imports single stranded DNA across the inner-membrane and into the cytoplasm of gram-negative bacteria (Pimentel and Zhang, 2018; Seitz and Blokesch, 2013), expression was verified by RT-PCR using primers specific for 'Ca. E. agilae' *comEC* and sequencing of the transcript from a protist cell fraction cDNA sample, prepared from 20 worker termite hindguts (Figure 7C). These data supported that these genes involved in this competence pathway are both conserved and expressed in the *Endomicrobium* symbionts of hindgut protists of *R. flavipes*.

The competence genes discussed above are involved in the translocation of single-stranded DNA across the inner membrane of gram-negative bacteria and subsequent recombination. Also present in the genomes of all five *Endomicrobium* species analyzed in this study are some genes which encode proteins that are homologous to components of Type IV pilus and *tad* locus/Type II secretion systems (T2SS), which may bind and import double-stranded DNA across the outer membrane (Ellison et al., 2018; Seitz and Blokesch, 2013). These may be components of systems which are in a state of decay in these *Endomicrobium* species, however '*Ca*. E. agilae' appeared to have retained a more complete *tad* locus compared to the other *Endomicrobium* species. In addition, all five genomes possessed a pre-pilin peptidase (PilD) as well as genes that encode one or more pilin types. A graphical summary of these findings along with the putative competence mechanism is provided as Supplemental Figure 6 and a list of found genes and their putative function can be found in Supplementary Table 3.

Discussion

Single protist cell metagenomics has enabled the assembly and recovery of genomes from several protist-associated bacterial symbionts in termites (Hongoh et al., 2008a, 2008b; Sato et al., 2009; Strassert et al., 2016; Utami et al., 2019). In this study, we present the near-complete draft genomes of several endosymbiotic *Endomicrobium* species, from three different protist types. These displayed differences in gene content with regards to several niche specific traits such as aerotolerance and carbon usage. These differences may highlight specific adaptations allowing for their persistence in protist hosts, acquired through HGT.

Differences in gene content related to both aerotolerance and carbon usage may reflect different selective pressures presented to each of these endosymbiont populations across different protist hosts. As '*Ca*. E. pyrsonymphae' resides in the cytoplasm of *P. vertens* which is usually found attached to the oxic gut wall, their retention of the complete aerotolerance operon (Batl) suggested that these intracellular bacteria encountered oxygen. In contrast, the other *Endomicrobium* species have lost at least one gene of this operon, which suggested that there may be relaxed selective pressure to maintain that operon either due to less exposure to oxygen or the acquisition of other adaptations to replace that function.

In addition to the *BatI* operon, the *Endomicrobium* genomes encoded other genes related to other oxidative stress responses, implying that even though *T. agilis* and *D*. species II are motile and reside in the anoxic hindgut, they still may encounter oxygen. This could be due to peristalsis, or mixing, which may temporarily disrupt the steep oxygen gradient in the termite hindgut by mixing the gut contents. Nevertheless, the presence of aerotolerance and oxidative stress response genes supported that these endosymbionts may detoxify their cytoplasm and maybe that of their protist hosts by neutralizing harmful oxygen species. Similar phenomena have been previously reported in sap-feeding insect hosts where the enhanced metabolic

activities in bacteriocytes generate oxidative stresses that are combated by increased expression of genes involved in oxidative stress responses in their endosymbionts (Pontes et al., 2008).

We noted that each *Endomicrobium* species had genes allowing for the usage of different carbon sources was an interesting finding. One hypothesis to explain these differences is that the carbon sources are provided to the endosymbionts by the protist hosts, as by-products (metabolites) of host-protist metabolism. For example, glucuronate may be present in the cytoplasm of *Trichonympha* spp. because they possessed the enzymes needed to cleave those monomers from certain polysaccharides found in wood whereas the other protists, *P. vertens* and *D*. species II, did not. This indirectly suggests that there may be specialization among the protist community with regards to polysaccharide hydrolysis in the hindgut of wood-feeding termites. An alternative hypothesis is that different symbionts, which co-colonize the same protist host, have partitioned the usage of metabolites within a given protist host. In support of this, the membrane-imbedded symbiont '*Ca*. Desulfovibrio trichonymphae' which co-colonized the same *Trichonympha* host as '*Ca*. E. trichonymphae' Rs-D17, used different carbon sources (malate and citrate) compared to its co-inhabitant whose genome also suggested that glucuronate and glucose-6-phosphate were carbon sources (Sato et al., 2009).

Endomicrobium species have acquired genes by HGT from several termite-associated bacterial taxa, thus these bacteria are susceptible to gene flow in the context of their associations with termites and their hindgut protists. Endosymbiotic lineages of *Endomicrobium* and their free-living relatives, possess many genes involved in DNA uptake, repair, and recombination (Supplemental Figure 5). We investigated genes which are known to be involved in natural transformation in other bacterial taxa. Our analyses showed that the genes *comEC*, *comF*, *ssB*, *drpA*, and *recA* are relatively conserved within the Elusimicrobia phylum and were expressed in

the endosymbiotic *Endomicrobium* species characterized in this study (Figure 7). Collectively these genes have been shown to be involved in the translocation of single stranded DNA across the inner membrane of gram-negative bacteria and homologous recombination (Seitz and Blokesch, 2013). The gene *comEC* in particular has an important function in this process as it encodes the DNA transporter (Pimentel and Zhang, 2018; Seitz and Blokesch, 2013). Using both transcriptome data, RT-PCR, and sequencing we were able to support that *Endomicrobium comEC* is expressed in the hindguts of termites. These data led us to the hypothesis that these endosymbionts may have the ability to become competent which would allow them to acquired genes from the termite gut community by HGT. It is still not clear how these organisms may transport DNA across their outer membranes. None of these *Endomicrobium* species possessed all the components of Type IV pili (T4P) and only '*Ca*. E. agilae' contained a near-complete *tad* locus/T2SS (Supplemental Figure 6 and Table 3). It is also puzzling as to why each of these genomes have retained *pilD* as well as genes that encode pilins. It may be that these pilins carry out some function in the periplasm or they could be are non-functional and are in the process of being lost. If competence is a common trait among the *Endomicrobium*, this could explain why these organisms appear susceptible to HGT, which may allow for relatively rapid adaptation to new or diverse niches. Since the hindgut protists phagocytize their foods, which includes wood and bacteria (Brune, 2014), this may be a route through which endosymbiotic *Endomicrobium* are exposed to exogenous DNA. However, it is worth noting that competence is not the only plausible avenue for gene acquisition which could also occur by bacteriophage transduction, conjugation, or other routes. Several lines of evidence that support that these endosymbionts are also susceptible to molecular parasites, such as bacteriophages and plasmids. Previous studies have reported that *Endomicrobium* species possessed several intact defense mechanisms to

combat molecular parasites such as CRISPR-Cas and restriction modification proteins (Izawa et al., 2016; Zheng et al., 2016a). The *Endomicrobium* spp. sequenced in this study also contained those defense systems. The complete genome sequence of a bacteriophage of an endosymbiont ('*Candidatus* Azobacteroides pseudotrichonymphae') of a termite hindgut protist has previously been published, supporting that susceptibility to phage infection is not limited to *Endomicrobium* endosymbionts (Pramono et al., 2017).

Our analysis of these *Endomicrobium* genomes obtained from single protist cell metagenomes, highlights several important differences across protist hosts which have led to us to interesting hypotheses that warrant further investigation. In each case, the major hurdle of testing precise hypotheses in these organisms is our current inability to culture them which restricts their experimental tractability. However, the use of -omics could still further our understanding of these protist-bacterial symbioses by focusing on focusing on the protist hosts' and their symbionts' gene, mRNA, and protein contents (Bhattacharya et al., 2011; Hamann et al., 2017; Karnkowska et al., 2016; Kolisko et al., 2014; Mangot et al., 2017; Vacek et al., 2018).

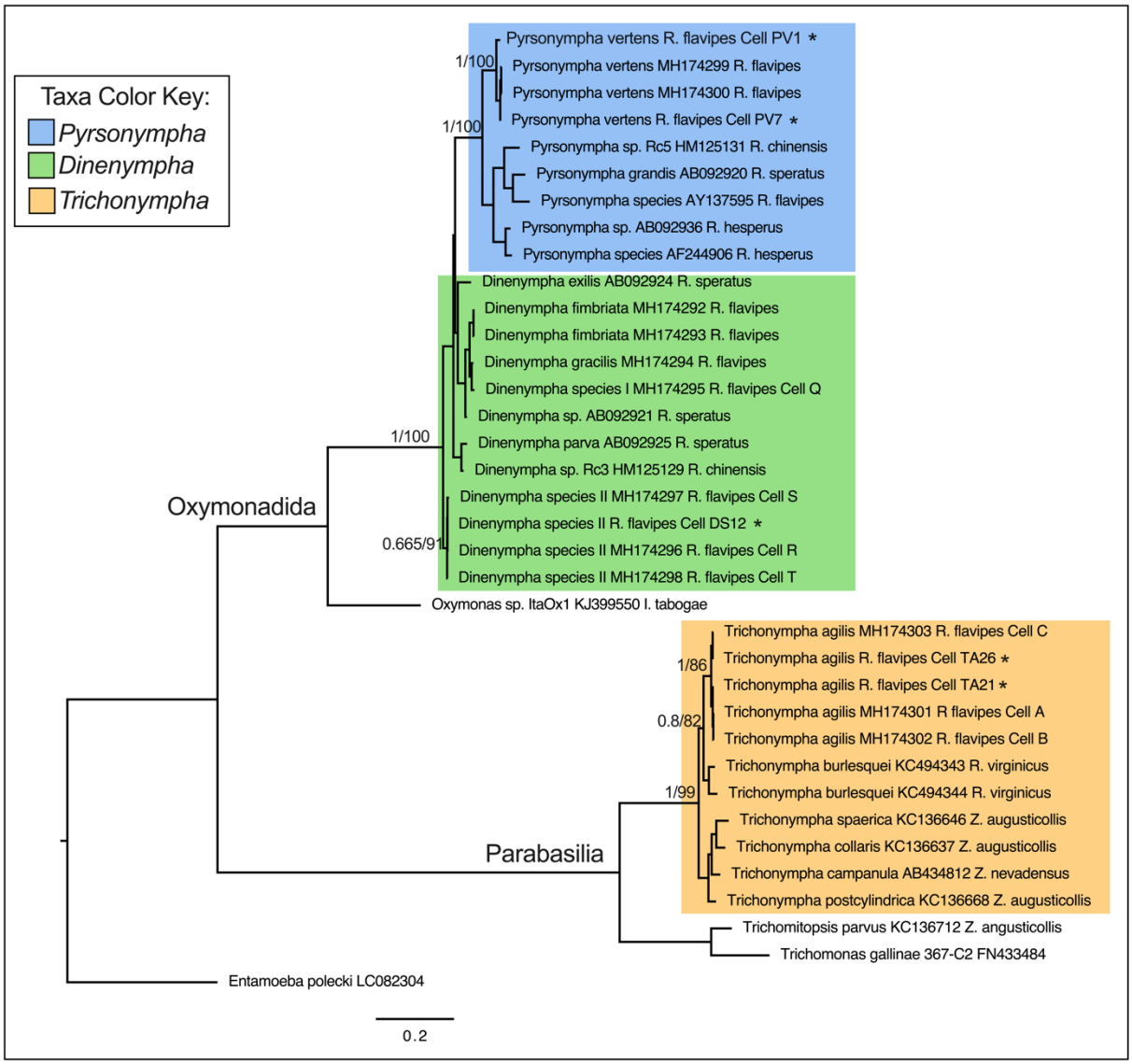


Figure 1 Protist 18S rRNA gene phylogeny. 18S rRNA genes were retrieved from single protist cell metagenome assemblies, aligned to references, and a Maximum likelihood (ML) phylogenetic tree was made using IQ-Tree using substitution model TIM2+G4. All 18S rRNA gene sequences obtained by this study (denoted by *) are shown grouped with their respective references. Branch support values represent the Bayesian posterior probability and Bootstrap support values respectively.

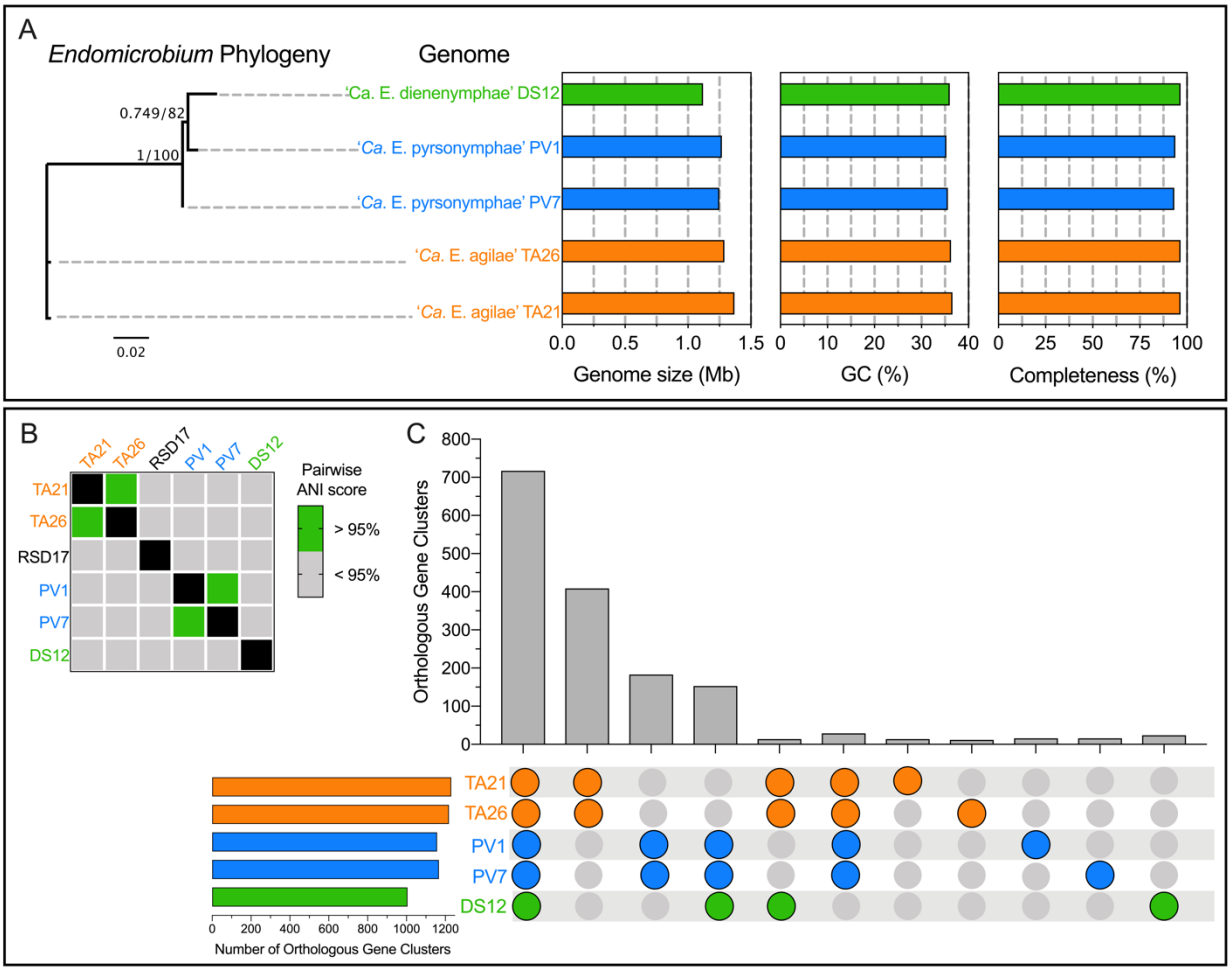


Figure 2 *Endomicrobium* draft genomes statistics, speciation, and shared gene content. (A) 16S rRNA gene Maximum likelihood tree (unrooted) of the three *Endomicrobium* species, genome sizes, percent G+C content, and estimated percent genome completeness. (B) Pairwise ANI scores of *Endomicrobium* genomes obtained by this study and a previously sequenced relative Rs-D17. (C) UpSet graph of the number of orthologous gene clusters (OGCs) of protein coding sequences within and across each of the *Endomicrobium* genomes.

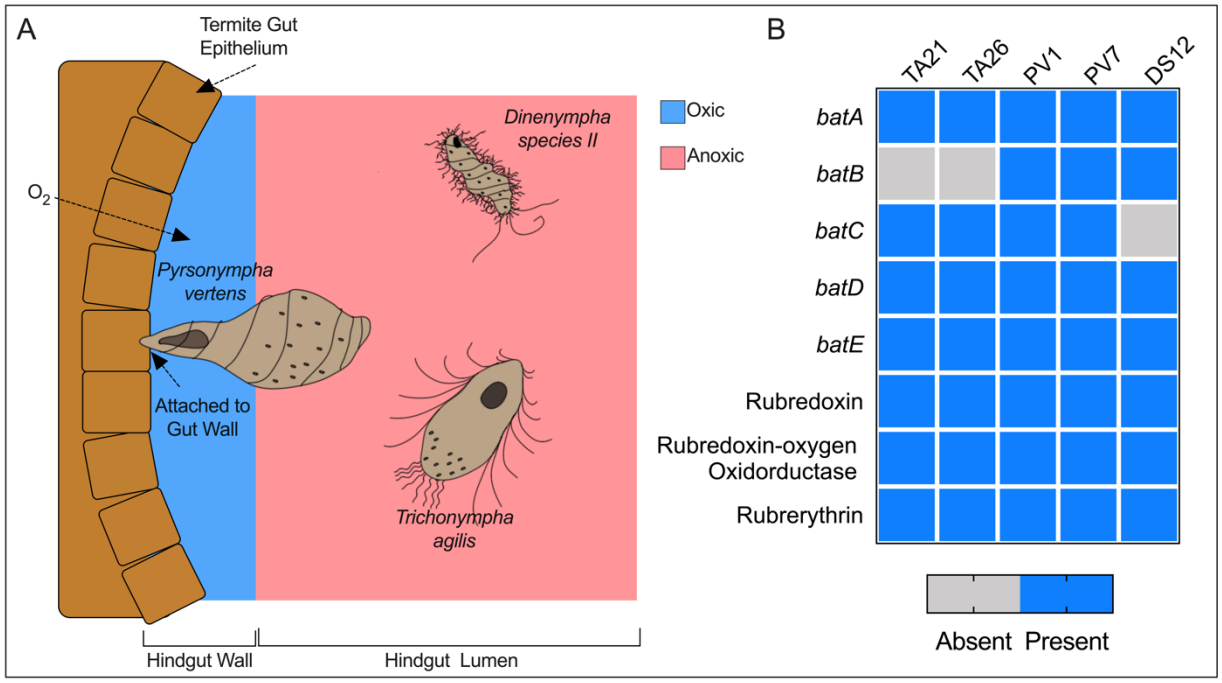


Figure 3 Differences in aerotolerance and oxidative stress response genes in the *Endomicrobium* genomes. (A) Diagram (not drawn to scale) of the relative positions of each protist host species within the termite hindgut and their inferred expose to oxic and anoxic regions. Oxygen diffuses across the hindgut wall, where *P. vertens* are usually found, whereas the hindgut lumen is anoxic (B) Presence/absence matrix of genes (rows) related to aerotolerance and oxidative stress responses found in the *Endomicrobium* genomes (columns).

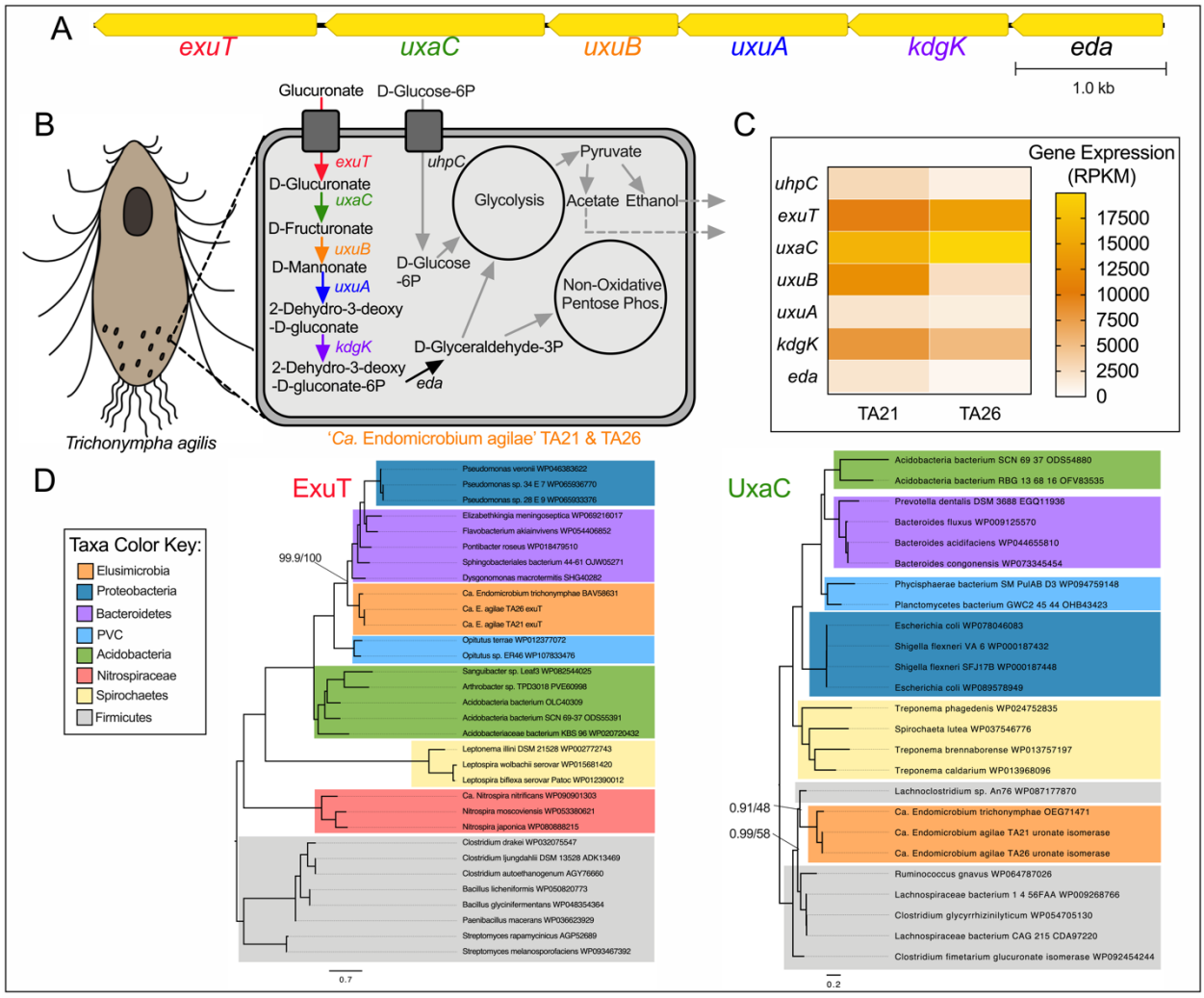


Figure 4 Carbon metabolism and HGT in '*Ca*. Endomicrobium agilae'. (A) Gene neighborhood of the genes involved in the metabolism of glucuronate in the '*Ca*. Endomicrobium agilae' TA21 and TA26 genomes. (B) Diagram of a protist host and an *Endomicrobium* cell showing the inferred metabolic conversions of carbon sources based on gene content data. (C) Gene expression data of genes of interest (rows) pertaining to carbon metabolism in '*Ca*. Endomicrobium agilae' TA21 and TA26 (columns). (D) Maximum likelihood phylogenetic trees of amino acid sequences of the transporter (ExuT, using substitution model LG+F+G4) and isomerase (UxaC, using substitution model LG+I+G4) in the

glucuronate metabolism pathway. Support values represent the Bayesian posterior probability and Bootstrap support values respectively.

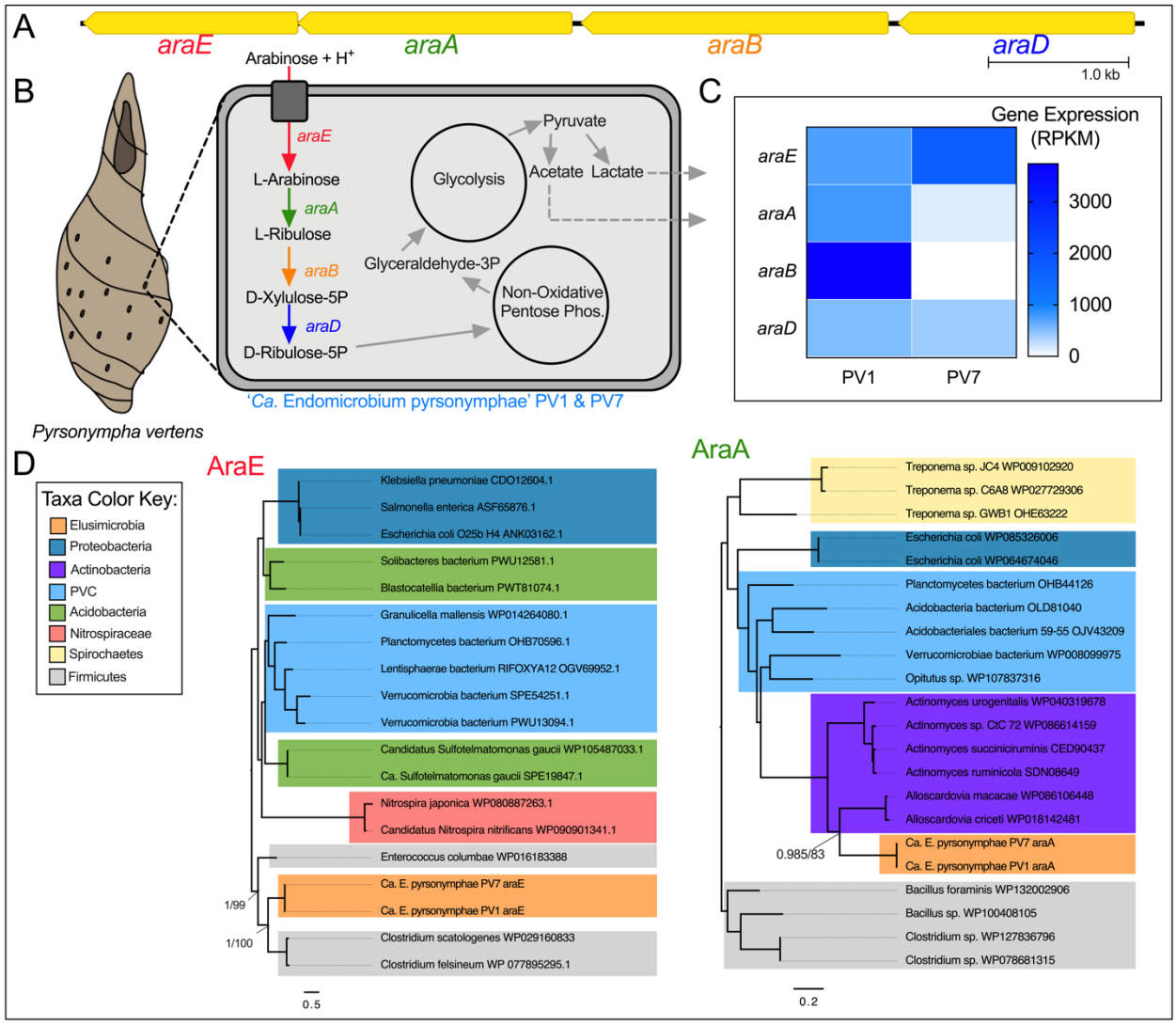


Figure 5 Carbon metabolism and HGT in '*Ca.* Endomicrobium pyrsonymphae'. (A) Gene neighborhood of the genes involved in the metabolism of arabinose in the '*Ca*. Endomicrobium pyrsonymphae' PV1 and PV7 genomes. (B) Diagram of a protist host and an *Endomicrobium* cell showing the inferred metabolic conversions of carbon sources based on gene content data. (C) Gene expression data of genes of interest (rows) pertaining to carbon metabolism in '*Ca*. Endomicrobium pyrsonymphae' PV1 and PV7 (columns). (D) Maximum likelihood phylogenetic trees of amino acid sequences from the transporter (AraE, using substitution model LG+F+G4) and isomerase (AraA, using substitution model LG+I+G4) in the arabinose

metabolism pathway. Support values represent the Bayesian posterior probability and Bootstrap support values respectively.

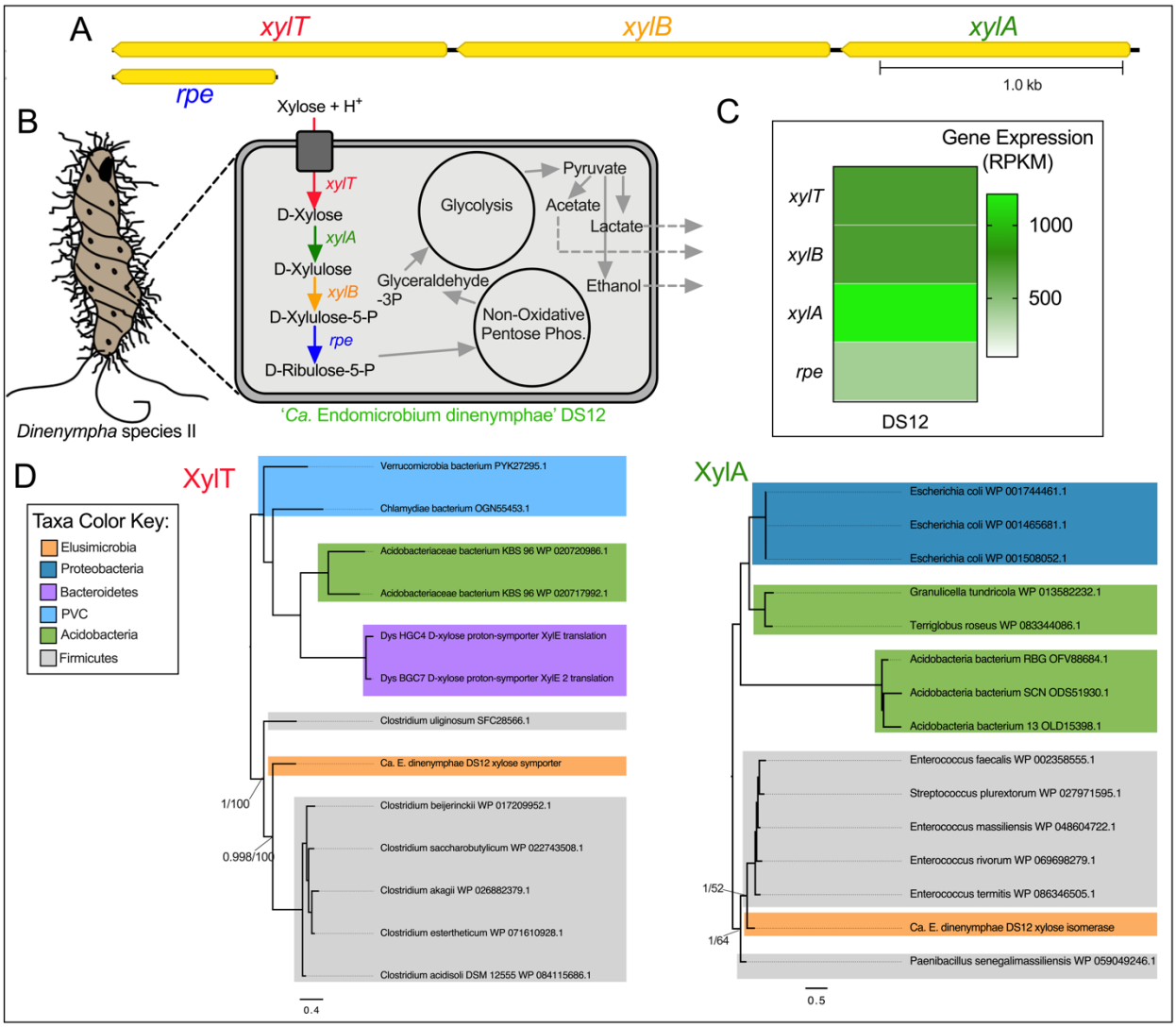


Figure 6 Carbon metabolism and HGT in '*Ca.* Endomicrobium dinenymphae'. (A) Gene neighborhood of the genes involved in the metabolism of xylose in the '*Ca*. Endomicrobium dinenymphae' DS12 genome. (B) Diagram of a protist host and an *Endomicrobium* cell showing the inferred metabolic conversions of carbon sources based on gene content data. (C) Gene expression data of genes of interest (rows) pertaining to carbon metabolism in '*Ca*. Endomicrobium dinenymphae' DS12 (column). (D) Maximum likelihood phylogenetic trees of amino acid sequences from the transporter (XylT, using substitution model LG+F+G4) and

isomerase (XylA, using substitution model LG+G4) in the xylose metabolism pathway. Support values represent the Bayesian posterior probability and Bootstrap support values respectively.

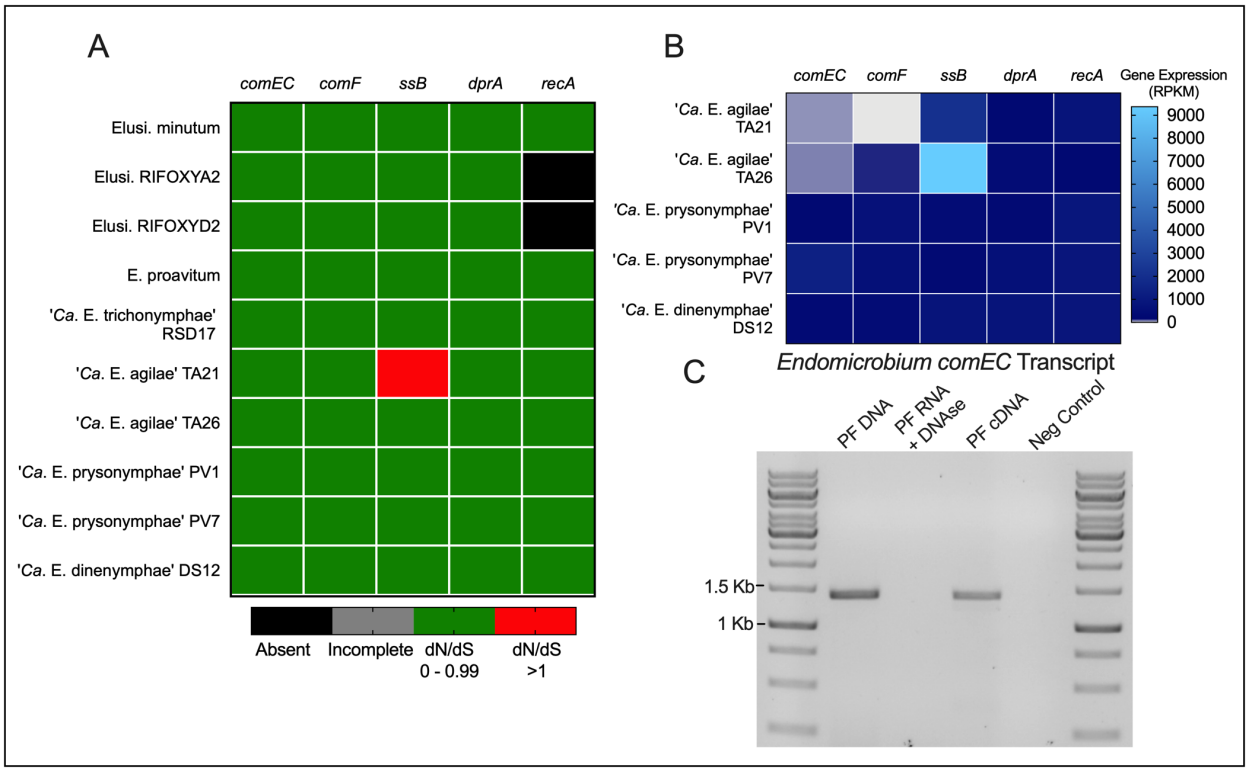
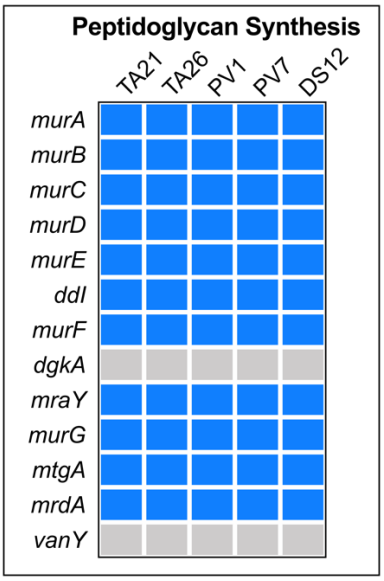


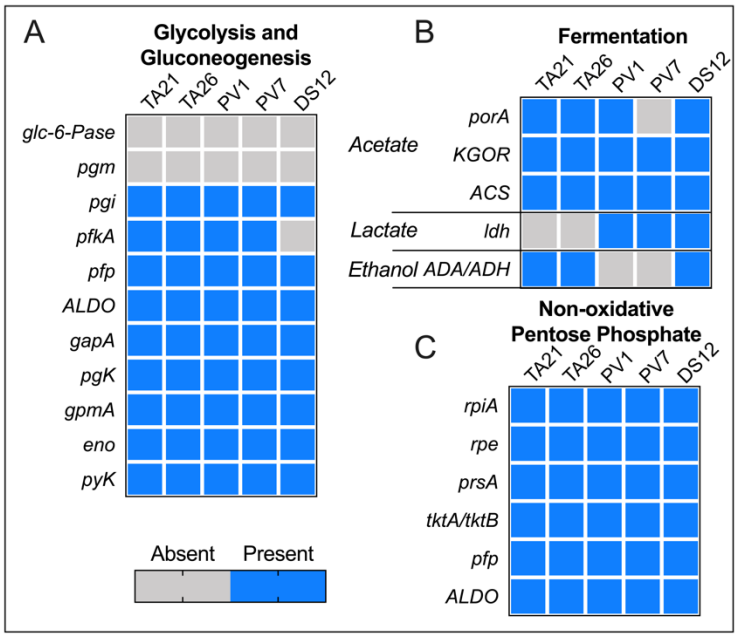
Figure 7 Analysis of genes involved in a putative competence pathway in *Endomicrobium* spp. (A) Heatmap showing the results of dN/dS analyses of genes involved in competence and recombination (columns) from *Endomicrobium* spp. and *Elusimicrobium* relatives (rows). (B) Gene expression data of those genes (columns) in the *Endomicrobium* spp. (rows) presented in this study. (C) RT-PCR gel image of *Endomicrobium comEC* transcript. Samples consisted of protist fraction (PF) DNA (positive control), PF RNA treated with DNase, PF cDNA, and molecular grade water (Negative control). Accession numbers for reference genomes used can be found in Supplementary Table 2.


A
Histidine
TA21 TA26 PV1* PV7 DS12
hisG
hisI
hisA
hisF
hisH
hisB
hisC
hisD
Arginine
TA21 TA26 PV1 PV7 DS12
argA
argB
argC
argD
argF
argG
argH
carA
carB
oat/argJ
Phenylalanine
TA21 TA26 PV1 PV7 DS12
aroG
aroB
aroD
aroE
aroK/aroL
aroA
aroC
pheA
AAT
Lysine
TA21 TA26 PV1 PV7 DS12
lysC
asd
dapA
dapB
dapC
dapH
dapD
dapE
patA
dapL
dapF
lysA
Valine
TA21 TA26 PV1 PV7 DS12
ilvN
ilvB
ilvC
ilvD
ilvE/BCA
Leucine
TA21 TA26 PV1 PV7 DS12
leuA
leuC
leuD
leuB
ilvE/BCA
Tryptophan
TA21 TA26 PV1 PV7 DS12
trpE
trpG
trpF
trpC
trpA
trpB
Isoleucine
TA21 TA26 PV1 PV7 DS12
ilvN
ilvB
ilvC
ilvD
BCA
Methionine
TA21 TA26 PV1 PV7 DS12
metA
metB
metC
metE
metT
Cysteine
TA21 TA26 PV1 PV7 DS12
cysE
cysK
CGL
Threonine
TA21 TA26 PV1 PV7 DS12
thrA
thrB/thrH
thrC
Absent Present
Glutamine and Glutamate
TA21 TA26 PV1 PV7 DS12
glnN
GOGAT
Glu
Symporter
Asparagine and Alanine
TA21 TA26 PV1 PV7 DS12
asnB
asnA
asdA
ALT
alaDH
Aspartic Acid and Tyrosine
TA21 TA26 PV1 PV7 DS12
GOT1
pheA2
tryA
hisC
Proline
TA21 TA26 PV1 PV7 DS12
proB
proA
proC
Serine and Glycine
TA21 TA26 PV1 PV7 DS12
Ser
Transporter
AGXT
glyA
Absent Present
B
Biotin Synthesis
TA21 TA26 PV1 PV7 DS12
Fatty Acids to Biotin
bioC
fabF
fabG
fabZ
fabI
bioH
bioI
Pimelate to Biotin
bioW
bioF
bioA
bioD
bioB
Absent Present
Pyruvate to Thiamine
Thiamine Synthesis
TA21 TA26 PV1 PV7 DS12
dxs
thiH
GO
iscS
thiF
thiI
thiG
tenI
thiE
TMK
Alk. Phos.
Aspartate to NAD+/NADP+
NAD+/NADP+ Synthesis
TA21 TA26 PV1 PV7 DS12
nadB
nadA
nadC
NMN deamidase
NADK
B12 to FMN and FAD
B12 Metabolism
TA21 TA26 PV1 PV7 DS12
btuC
RFK/
FADS
Lipoic Acid Synthesis
TA21 TA26 PV1 PV7 DS12
lipA
lipB

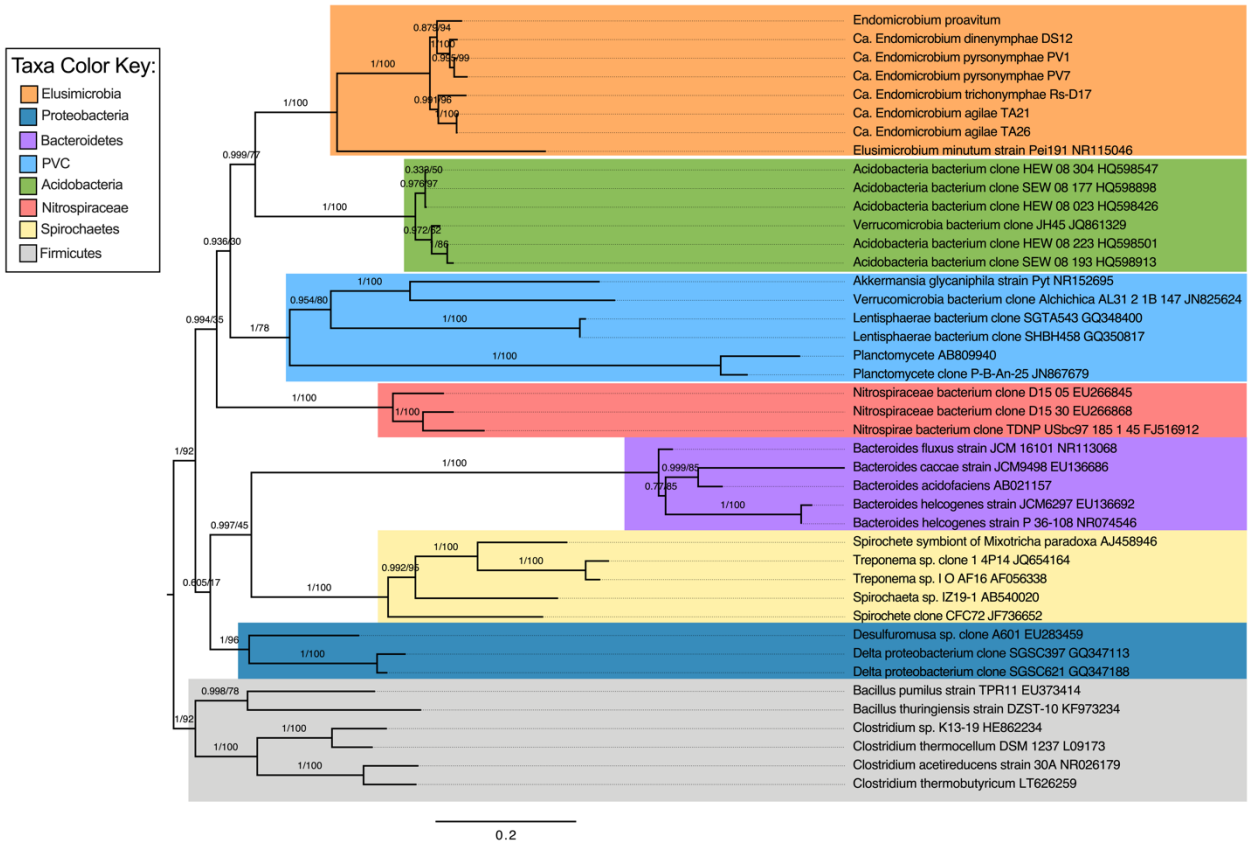
Supplemental Figure 1 Gene retention of biosynthesis pathways for amino acids and vitamins in *Endomicrobium* species. (A) Gene content of genes involved in the biosynthesis of amino acids and (B) vitamins. (*) Sample PV1 WGA reads were mapped to the PV7 assembly to assess the presence of the histidine biosynthesis genes based on mapping coverage.



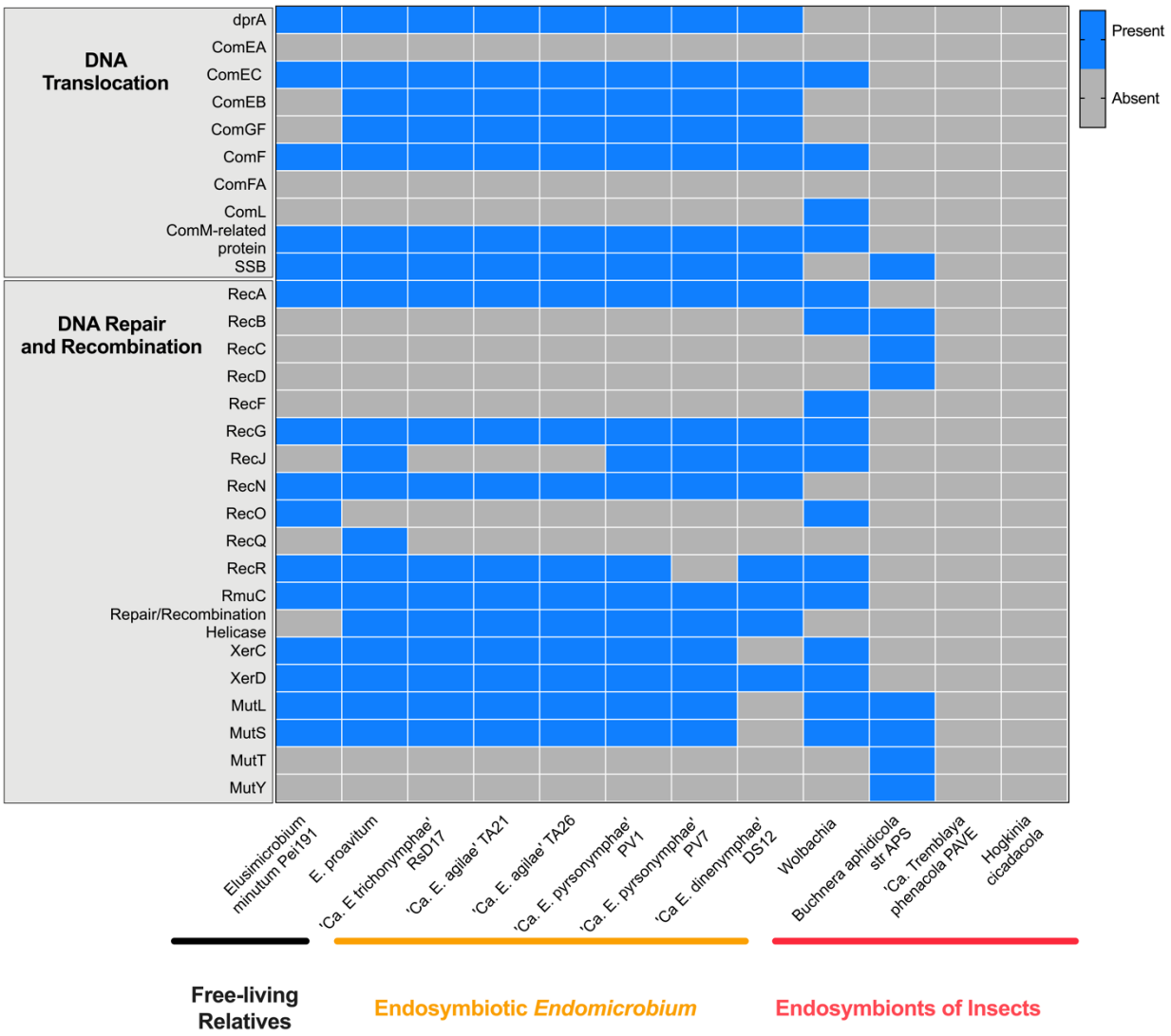
Supplemental Figure 2 Genes related to peptidoglycan biosynthesis. Genes involved in the biosynthesis of peptidoglycan are displayed as rows and their presence or absence is noted for each genome (columns). All five *Endomicrobium* genomes possessed an incomplete peptidoglycan synthesis pathway and were missing the same two genes (*dgkA* and *vanY*).



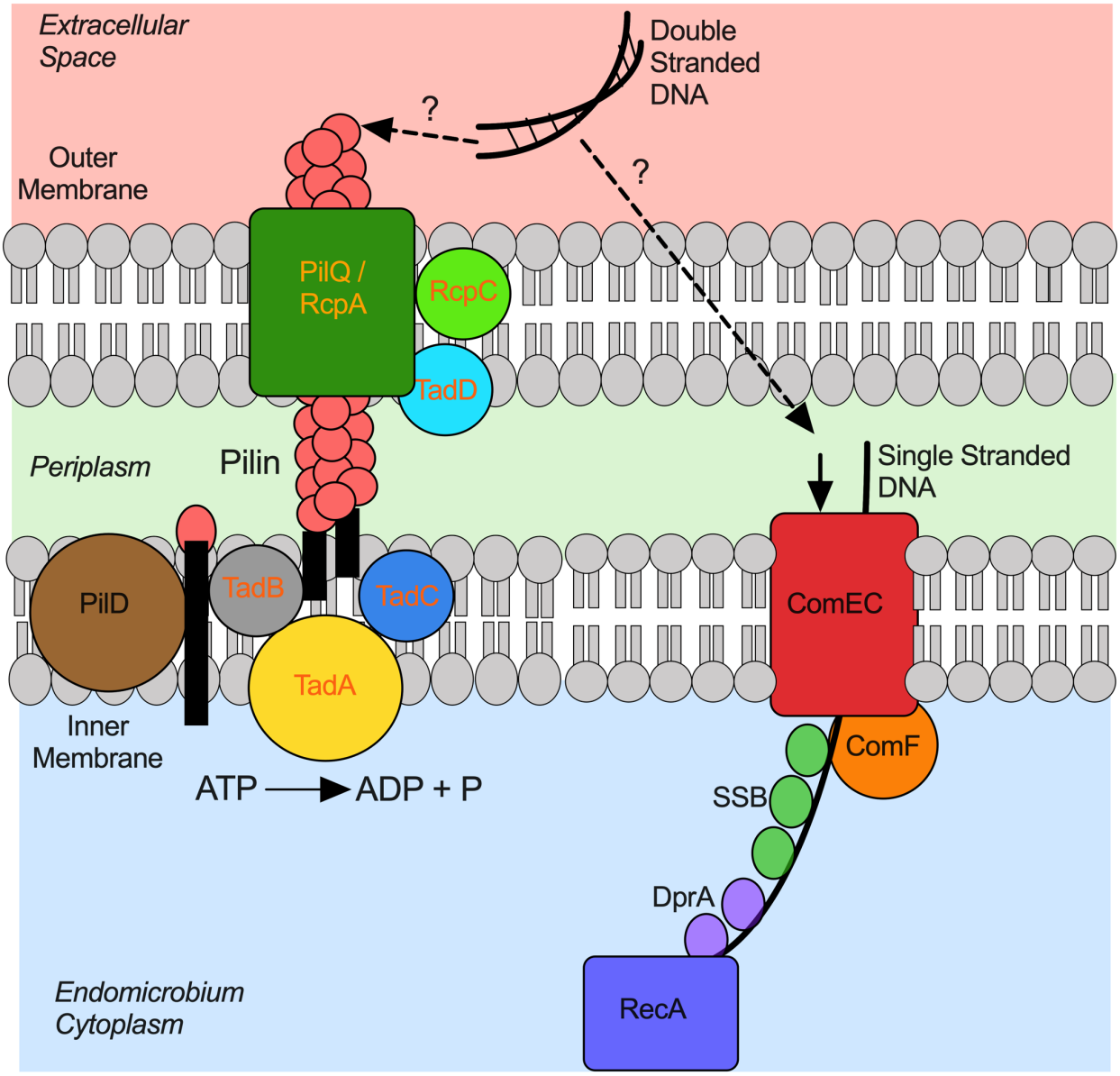
Supplemental Figure 3 Genes involved in central carbon metabolism. Genes (rows) involved in (A) glycolysis and gluconeogenesis, (B) fermentation, and (C) non-oxidative pentose phosphate in the *Endomicrobium* genomes (columns). The *Endomicrobium* have retained nearly identical gene content for genes involved in glycolysis, gluconeogenesis, and non-oxidative pentose phosphate. However, these genomes differ with regards to their fermentation of pyruvate.



Supplemental Figure 4 16S rRNA phylogeny of Elusimicrobia with respect to other phyla. Maximum likelihood phylogenetic trees of 16S rRNA which was used a marker-gene to establish an organismal phylogeny of the Elusimicrobia phylum. This phylogeny was used to determine incongruence between the 16S rRNA gene and other genes of interest that may have been acquired via HGT by the *Endomicrobium* species.



Supplemental Figure 5 Presence or absence of genes related to DNA translocation, repair, and recombination in *Endomicrobium* spp. compared to relatives and other endosymbionts. Presence/absence matrix of genes involved in DNA translocation, repair, and recombination were found in the *Endomicrobium* genomes and their presence or absence was investigated across free-living relative and other endosymbionts. The *Endomicrobium* and their relatives were enriched in genes related to those processes compared to other endosymbionts. Accession numbers for reference genomes used can be found in supplementary table 2.



Supplemental Figure 6 Graphical summary of a putative competence pathway and proteins involved in pilus assembly in *Endomicrobium* species. Proteins shared by all *Endomicrobium* species are in black font while those that are only retained in '*Ca*. E. agilae' are colored in orange. All *Endomicrobium* possessed the pre-pilin peptidase (PilD) and one or more genes that encode pilins. '*Ca*. E. agilae' possessed a near-complete *tad* locus/ T2SS as well the T4P secretin (PilQ).

Supplemental Table 1 Reference genomes and other sequences used in this study to filter contaminating genomic reads from single protist cell (meta)genome samples.

Genome, Vector, or Reference Dataset	Accession number
Malassezia globosa CBS 7966	AAYY01000001
Sinorhizobium meliloti 1021	NC003047
Propionibacterium acnes Strain PA_15_1_R1	CP012355
Propionibacterium acnes strain PA_21_1_L1	CP012351
Escherichia coli UMN026	NC011751
hg19_main_mask_ribo_animal_allplant_allfungus.fa.gz*	NA
Gateway vector pB4GWnY	AB830559

* Reference dataset can be found here: https://drive.google.com/file/d/0B3llHR93L14wd0pSSnFULUlhcUk/edit?usp=sharing

Supplemental Table 2 Reference genomes used in this study for gene content comparisons and/or dN/dS analyses.

Genome	Accession number
Elusimicrobium minutum Pei191	NC010644
Endomicrobium proavitum strain Rsa215	CP009498
Candidatus Endomicrobium trichonymphae Rs-D17	NC020419
Elusimicrobia bacterium RIFOXYA2	MGVD00000000
Elusimicrobia bacterium RIFOXYB2	MGVO00000000
Candidatus Hodgkinia cicadicola	CM008689
Candidatus Tremblaya phenacola PAVE	CP003982
Wolbachia species	NC002978
Buchnera aphidicola str. APS	NC002528

Supplemental Table 3 Proteins involved in pilus assembly in the *Endomicrobium* genomes.

Function	TA21	TA26	PV1	PV7	DS12
Pilins	PilA	PilA	PilA	PilA	PilA
	TadE	TadE	PilE	PilE	-
	Flp	Flp	-	-	-
Secretins	PilQ/RcpA	PilQ/RcpA	-	-	-
Platform proteins	TadB	TaB	-	-	-
	TadC	TadC	-	-	TadC (partial)
Assembly ATPases	TadA	TadA	-	-	-
Prepilin peptidases	PilD	PilD	PilD	PilD	PilD (partial)
Unknown	TadD	TadD	TadD	TadD	-
	RcpC	RcpC	-	-	-

Chapter Four

Analysis of Ectosymbiotic *Treponema* Communities Supports Their Role in Essential Supportive and Hydrolytic Functions in Wood-Feeding Termites[+]

Contribution from other researchers

Dr. Jacquelynn Benjamino contributed to this chapter. She performed whole genome and transcriptome amplification as well as prepared sequencing libraries.

[+] In preparations for submission, Stephens, Benjamino, Graf, and Gage.

Abstract

The ability of wood-feeding termites to live exclusively on nutrient poor-lignocellulose relies on the metabolic actives on their hindgut microbiota to aid in hydrolyzing the polysaccharides found in wood, recycling fermentation end products, and assimilating essential nitrogenous compounds. In this study we use a combination of assembly and read-based omic analyses to investigate whether ectosymbiotic communities of *Treponema*, which are associated with hindgut protists, contribute to these essential functions. Our results supported that ectosymbionts of *Trichonympha* and *Dinenympha* protists possessed and expressed genes which encoded glycoside hydrolases, the Wood-Ljungdahl pathway, and nitrogenase. Collectively these *Treponema* species likely make significant contributions to the necessary metabolic conversions of carbon and nitrogen substrates which enable their hosts to thrive on a nutrient-poor diet.

Introduction

The symbiotic protists that inhabit the hindguts of wood-feeding termites are often colonized by ectosymbiotic communities of bacteria (Ohkuma, 2008). The composition of these communities can vary across different termites but in *Reticulitermes* termites, their hindgut protists are colonized by ectosymbionts which include *Treponema* species and '*Candidatus* Symbiothrix'(Noda et al., 2003, 2006; Yuki et al., 2015). Since termite protists cannot currently be cultivated, we are limited in our understanding of their associations with ectosymbionts. Confounding this issue is the fact that *Treponema* dominate the bacterial hindgut community of wood-feeding termites like *Reticulitermes flavipes* (Benjamino and Graf, 2016), and are commonly found to be free-living in the hindgut fluid. In addition, ectosymbiotic communities of *Treponema* have been shown to exhibit a high level of diversity on the surface of protist hosts

(Stephens and Gage, 2018), and include members of at least two phylogenetic clusters (referred to as Termite *Treponema* Clusters I & II) (Iida et al., 2000). Because of their large populations and diversity across different niches within the hindguts of wood-feeding termites, it is challenging to assign certain functions to any one group of these *Treponema*.

There are several hypotheses pertaining to the functions of the ectosymbiotic *Treponema* which stem from the observed metabolisms of cultivated relatives which have been isolated from termite hosts. These functions include the ability to fix both carbon and nitrogen, which are regarded as essential functions in the hindguts of wood-feeding termites (Brune, 2014; Lilburn et al., 2001; Ohkuma et al., 1996; Pester and Brune, 2006). While there are several lines of evidence supporting that members of the *Treponema* do participate in those functions in termites(Lilburn et al., 2001; Pester and Brune, 2006), it has not yet been determined if this extends to the ectosymbiotic *Treponema*.

With regards to carbon fixation, previous studies have supported that termite-associated *Treponema* expressed a marker gene (*fhs*; formyltetrahydrofolate synthetase) for reductive acetogenesis using the Wood-Ljungdahl pathway in *R. flavipes* (Pester and Brune, 2006). In doing so, they convert CO_2, which is a fermentation end product, into the acetate which is the primary carbon source of the termite hosts (Brune, 2014; Pester and Brune, 2006). This function is thought to be essential since it provides a necessary sink for fermentation end products and allows for the continual microbial fermentation of liberated sugars from the polysaccharides found in wood. In addition, the cultivated isolates *Treponema primitia* ZAS-1 and ZAS-2 possessed all genes in the Wood-Ljungdahl pathway and were observed to fix CO_2 and produce acetate *in vitro*, supporting that reductive acetogenesis may be a common metabolic feature of termite associated *Treponema* species (Graber and Breznak, 2004). Only recently has there been

direct evidence which supports that at least some protist-associated *Treponema* may be also acetogens. A recent analysis of a draft genome of an ectosymbiotic *Treponema* species' genome supported that it too possessed all the necessary genes of the Wood-Ljungdahl pathway (Utami et al., 2019).

Nitrogen is thought to be limiting in the hindgut of wood-feeding termites since their food source (lignocellulose) is deficient in essential nitrogenous compounds, and lignin the major source of nitrogen in wood, is largely undigested by termites (Brune, 2014). Instead, nitrogen enters into the hindguts of termites by diffusing across the termite's tissues as N_2 gas and is converted into ammonia by nitrogen fixing bacteria (Brune, 2014; Ohkuma et al., 1996). Previously, studies have concluded that nitrogen fixation is carried out by a broad group of bacterial taxa within the hindguts of wood-feeding termites (Ohkuma et al., 1996). Several lines of evidence support that *Treponema* are among that group (Lilburn et al., 2001). The termite *Treponema* isolates ZAS-1, ZAS-2, and ZAS-9 all possessed *nifH* homologs in their genomes and were able to fix N_2 *in vivo*, which was measured using an acetylene reduction assay (Lilburn et al., 2001). However, it is not currently known if the ectosymbiotic communities of *Treponema* are able to fix nitrogen.

A previous a genomic analysis of '*Ca*. Symbiothrix' which is a member of the ectosymbiotic community of some *Dinenympha* spp. protist hosts in *Reticulitermes* spp. (Yuki et al., 2015) concluded that this particular ectosymbiont's genome was enriched in glucoside hydrolase family (GHF) enzymes, which suggested that it participates in polysaccharide hydrolysis (Yuki et al., 2015). GHF enzymes may also be encoded by other ectosymbionts such as the *Treponema*, however that has not yet been investigated. The possibility of polysaccharide hydrolysis by the ectosymbionts of hindgut protists offers a new insight into these intimate

symbioses which have traditionally been studied through the lens of the hypothesis of metabolite exchange.

Both *Treponema* and '*Ca*. Symbiothrix' ectosymbionts are attached to the surface of their protist hosts by one cell pole. The interface of the binding sites has been shown to include electron dense materials which connect the outer membrane of the ectosymbionts to the plasma membrane of the protist hosts (Radek and Tischendorf, 1999; Yuki et al., 2015). A previous study investigated the attachments of ectosymbiotic spirochetes and rod-like bacteria to two protist species which inhabited the hindgut of *Mastotermes darwiniesis* (Radek and Nitsch, 2007). Their analyses supported that the ectosymbionts were attached to their protist hosts by protein mediated mechanism which was disrupted by treatments with urea (Radek and Nitsch, 2007). This was further supported by freeze-fracture experiments which revealed integral membrane proteins that displayed periodicity in their arrangements at attachment interface (Radek and Nitsch, 2007). Although it is not currently known which proteins mediated these attachments, the genome of the ectosymbiont '*Ca*. Symbiothrix' did contain genes which could encode proteins similar to known adhesion/binding proteins (Yuki et al., 2015). These candidate binding proteins included immunoglobulin-like proteins, fibronectin type III domain containing proteins, and ankyrin-like repeat containing proteins (Yuki et al., 2015).

In this study we used single isolated protist cells isolated from the hindgut of *R. flavipes*, as templates for whole genome and whole transcriptome amplification and sequencing to produce both metagenomes and metatranscriptomes of the bacterial symbionts attached to each protist cell. We focus our analyses on the ectosymbiotic *Treponema* communities, which are currently only poorly understood with regards to their contributions to essential hindgut functions and the molecular mechanisms of their attachment to host cells. Using contig binning,

we recovered a *Treponema* genomic bin corresponding to a subset of the *Treponema* community associated with *Dinenympha* species II. This was used to investigate gene content related to reductive acetogenesis, nitrogenase, GHF enzymes, and putative binding proteins. We further investigated the transcription of these genes at the community level using metatranscriptomes. Collectively, our assembly and read-based analyses of the ectosymbiotic communities supported the hypothesis that ectosymbiotic *Treponema* participate in reductive acetogenesis, nitrogenase activity, and polysaccharide hydrolysis. In addition, we show that the interface of attachment is mediated by electron dense materials which can be cleaved during proteinase treatment, supporting that ectosymbionts are attached to their protist host by protein-mediated attachment as seen in previous studies (Radek and Nitsch, 2007). We also propose certain candidate proteins that may be involved in the attachment which can be targeted for future studies.

Methods

Termite Collection and Identification

Termites were collected from cardboard traps which were placed under actively infested logs in Mansfield CT USA. The species identity was confirmed as previously described, by PCR amplification and sequencing of the mitochondrial cytochrome II oxidase gene (Stephens and Gage, 2018).

Protist Isolation and Whole-Genome and Whole-Transcriptome Amplification

Single protist cells were isolated from termite hosts and frozen as previously described. Briefly, worker termites were brought into a an anaerobic chamber and their hindguts were dissected and ruptured into ice-cold Trager's solution U (TU) (Trager, 1934). Protist cells were washed as previously described using ice cold TU followed by dilution in ice-cold TU. Single protist cells were isolated and washed three times using an Eppendorf Cell Tran Vario equipped with a hand-drawn glass capillary tube. In an attempt to increase coverage of bacterial reads, for one sample (TA.enuc) the nucleus of the protist was removed by gentle lysis by exposure to 5 mM KOH and 0.025% Triton-X to rupture the protist cell followed by micromanipulation. Following isolation, single protist cells were placed in molecular grade phosphate buffered saline (PBS), flash frozen on dry ice, and immediately stored at -80 degrees Celsius.

Whole genome and whole transcriptome amplification were performed as previously described. The metagenome (DNA) and metatranscriptome (cDNA) from the protist cell samples were amplified between 12-24 hours after isolation. Cell lysis and amplification was performed using the Repli-g WGA/WTA kit (Qiagen). Cells were lysed using a Qiagen lysis buffer and incubation step, immediately followed by incubation on ice. Aliquots from the same lysed cell were separated and used in simultaneous whole genome amplification (WGA) and whole transcriptome amplification (WTA). The process was carried out per the kit's standard protocol with exception of the addition of random hexamer primers which were used to amplify DNA and cDNA.

Library Preparation and Sequencing

DNA and cDNA were sheared using a Covaris M220 ultra-sonicator™ according to standard protocol. WGA samples were sheared to a 550 bp insert size using 200 ng of DNA and WTA samples were sheared to a 350 bp insert size using 100 ng of cDNA. Sequencing libraries were prepared using the TruSeq Nano DNA Library Prep kit from Illumina® according to standard protocol. Each sample was prepared with a forward and reverse barcode for the ability to add all samples on the same sequencing run. The samples were sequenced using an Illumina® NextSeq 1 x 150 Mid Output run and two NextSeq 1 x 150 High Output runs. For sample TA.enuc all sequencing was carried out on an Illumina® MiSeq (2 x 250 run).

Metagenome Assembly, Genomic Binning, and Annotation

(Meta)genomic and (meta)transcriptomic reads were filtered and trimmed as previously described and the (meta)genomics reads were assembled as previously described. Reads were preprocessed before assembly with a BBmap (Bushnell, 2014) workflow. Reads were filtered for contaminating sequences by mapping reads to reference genomes of potential contamination sources such as human DNA, human associated microbiota, and organisms commonly used in our research laboratories. A list of references genomes used for contamination filtering is provided as a Supplementary Table 1. Using the "bbduck.sh script", adaptor sequences were trimmed from reads, and last base pair of 151 or 251 bp reads were removed using the "ftm" command. Reads were trimmed at both ends using the quality score of Q15 as a cutoff. Homopolymers were removed by setting an entropy cutoff of 0.2, a max G+C cutoff of 90%, and by removing reads which possessed stretches of G's equal to or greater than 25 bases long. In addition, reads which were below a minimum average quality of Q15 and/or 50 bases long were

removed. Genomic reads were then normalized to a minimum coverage of 2X and a maximum coverage of 50X and then deduplicated using BBnorm. Genomic reads were then assembled using the A5 assembly pipeline (Coil et al., 2015) on the KBase web server (Arkin et al., 2018).

Metagenomic assemblies of single protist host cells and their bacterial symbionts were binned as using the program MAxBin2 (Laczny et al., 2015). One genomic bin of interest, which contained a single 16S rRNA gene belonging to a *Treponema* species, was selected from a sample *Dinenympha* species II cell called "DS12" for further analysis (referred to as *Treponema* bin DS12). The program CheckM (Parks et al., 2015) was used to asses completion which scored this bin as being 100% complete but also found that every marker gene was duplicated. This along with the observation that this bin was twice as large as other termite-*Treponema* genomes Table 2) led us to suspect that this bin likely represented two genomes. This genomic bin was then annotated on the RAST Server (Aziz et al., 2008) using a customized RASTtk (Brettin et al., 2015) workflow with additional options selected to call insertion sequences and prophages.

To investigate the presence and diversity of GHF enzymes encoded by *Treponema* bin DS12, the meta server dbCAN2 (Zhang et al., 2018) was used. Putative genes which encoded GHF enzymes were considered only if bdCAN2 reported a hit using two or more methods. This analysis was repeated using the genomes of the cultivated relatives ZAS-1, ZAS-1, and ZAS-9 and the GHF content of these termite *Treponema* were compared to one another. GHF enzymes were grouped into putative functional categories based on their reported substrate activities in the Carbohydrate Active Enzymes database (CAZy; http://www.cazy.org/) (Lombard et al., 2014).

Phylogenetic Analyses of Ribosomal rRNA Genes

The 16S rRNA gene of *Treponema* bin DS12 was aligned to references sequences belonging to other termite associated *Treponema* using MUSCLE (Edgar, 2004) and a phylogenetic tree was generated with a Maximum likelihood (ML) method using the program IQ-Tree (Nguyen et al., 2015) with model testing. Protist 18S rRNA gene sequences were retrieved from metagenome assemblies, aligned to reference sequences using MUSCLE and a ML phylogenetic tree was made using the IQ-Tree with model testing.

Read-based Metatranscriptome Mapping and Analysis

(Meta)transcriptome reads which had been trimmed and filtered as described above were used in DIAMOND (Buchfink et al., 2014) Blast search against the RefSeq (Pruitt et al., 2007) release 78 reference database using an e-value cutoff of 0.001. The resulting DIAMOND databases files were then mapped to KEGG (Kanehisa and Goto, 2000) and InterPro2GO (Mitchell et al., 2015) databases using MEGAN (Huson et al., 2016) which performs taxonomic and functional grouping of reads. Reads belonging to the bacterial phylum Spirochaetes were retrieved from each single protist single-cell metatranscriptome and then normalized using MEGAN. The number of normalized reads mapping to genes and metabolic pathways of interests were compared between samples. For genes involved in the Wood-Ljungdahl pathway and those that encoded nitrogenase, KEGG annotations were used. For genes classified as encoding GHF enzymes, InterPro2GO annotations were used. Statistical significance in the differences of the proportions of mapped reads were tested using two-tailed Mann-Whitney tests in GraphPad Prism version 8.

Ectosymbiont Detachment Assays

To test perturbations which detach ectosymbiotic bacteria from their protist hosts, protist cells were collected and washed as described above. The cells were then fixed with in 2% glutaraldehyde in TU for 1 hour at room temperature (RT) or kept alive prior to detachment assays. Using modified methods from a previous study (Radek and Tischendorf, 1999), live and fixed cells were treated with chemicals which disrupt certain molecular interactions. Treatments included 3M NaCl, 1M $CaCl_2$, and 10mM ethylenediaminetetraacetic acid (ETDA) in an attempt to disrupt ionic interactions; 0.1 and 0.3% Triton-X for hydrophobic interactions, and 0.1 and 0.2mg/mL of proteinase K for protein interactions. Following exposure to these treatments, detachment was followed by light microscopy until the protist cells lysed or began to alter in morphology, which generally was observed within five to ten minutes. Treatments which caused ectosymbionts to detach from their protist hosts were repeated and the samples were prepared for electron microscopy.

Transmission and Scanning Electron Microscopy

Scanning electron microscopy (SEM) was used to investigate the site of ectosymbiont attachment following treatment of proteinase K. The hindguts of worker termites were ruptured in 500µl of Trager's solution U (TU) and then centrifuged for 3 minutes at 5,000 rpms. The cells were then washed with another 500µl TU then split into two groups. One group was treated with proteinase K (0.1mg/ml) for 30 minutes while the other group was exposed to a negative control buffer which was used to re-constitute the proteinase K stock. Protist cells were then collected by low spin centrifugation as described above and fixed in 2% glutaraldehyde in TU (pH 7) at RT in

an anaerobic chamber as described above. Fixed cells were prepared for SEM as previously described (Stephens and Gage, 2018). Samples were deposited onto poly-L-lysine coated silicon wafer chips (Prod No. 16008, Ted Pella Inc.), washed with 80 mM sodium cacodylate buffer (pH 7), and post-fixed in 2% osmium tetroxide at RT for 1 hour. The cells were rinsed twice for 5 minutes in distilled water then dehydrated in serial concentrations of ethanol (30%, 50%, 70%, 95%, 100%, 5 min each), and critical point dried (931GL, Tousimis). Samples were then mounted on SEM stubs using silver paint, sputter coated with palladium (E5100, Polaron), and examined using a scanning electron microscope (Nova NanoSEM 450, FEI).

Transmission electron microscopy (TEM) was carried out on negative control samples to visualize the attachment site of ectosymbionts on protist hosts. Fixed and washed cells were spotted onto a poly L-lysine coated piece of Aclar and incubated at RT for 1-2 hours. The cells were then rinsed once in TU then washed three times in 80mM cacodylate buffer (CB) for 5 minutes each. Cells were post fixed in 80mM CB with 2% osmium tetroxide (OSO_4) in a dark container for 1 hour at RT.

The samples were then subjected to a dehydration series which consisted of 10-minute incubations of 30%, 50%, 70%, 95%, 100% x 2 of ethanol and then 100% acetone x 2. Samples were then placed in a 1:1 mixture of resin (Glauert medium media) and 100% acetone for 1 hour at RT with slight shaking. This step was then repeated with a 2:1 mixture of resin to acetone then 100% resin. The samples were then sandwiched between layers of Aclar with cutout areas roughly the size of the samples and then polymerized at 65 degrees Celsius for 48 hours. The polymerized resin was then peeled from the Aclar and visually inspected for the presence of preserved protist cells by light microscopy. This produced a monolayer of protist cells embedded in resin.

Areas in which protist cells were present were trimmed and glued to polymerized blocks of resin. Block faces were then trimmed to approximately 3mm x 3mm and then sectioned with an ultra-diamond knife producing 0.1µm ultrathin sections. Sections were placed on thin bar grids and left to air dry. Samples were then stained with 2% uranyl acetate for 8 minutes followed by washes in diH_2O then stained with Sato's lead citrate for 2 minutes, washed with diH_2O, and then air dried. Samples were then viewed by TEM using the FEI Tecnai G2 Spirit BioTWIN microscope.

Results

Gene content and expression of genes related to essential hindgut functions in ectosymbiotic *Treponema* communities

Single protist cell metagenome and metatranscriptome samples served as templates for both assembly and read-based approaches to investigate the communities of ectosymbiotic *Treponema*. These communities were surveyed across individual *T. agilis* and *Dinenympha* species protist hosts (Supplementary Figure 1) which co-inhabit the hindgut of *R. flavipes*. We retrieved the 18S rRNA gene from the metagenome assemblies for most of the protist cells used in this analysis (Supplementary Figure 1). One genomic bin was recovered from a single protist cell sample (referred to as *Treponema* bin DS12). This genomic bin contained a single 16S rRNA gene which grouped with other termite *Treponema* belonging to cluster I (Figure 1). This cluster also included the cultivated strains *T. primitia* ZAS-1 and ZAS-2, which are acetogens (Figure 1). Since the size of this genomic bin is approximately twice the size of previously sequenced relatives and most genes appeared to be duplicated, this genomic bin likely

represented two *Treponema* genomes (Table 2). In this study, we refer to this genomic bin as a representation of a subset of the ectosymbiotic community of *Treponema* associated with *Dinenympha* species II cell DS12.

The diversity of GHF enzymes was investigated in *Treponema* bin DS12 and compared to the GHF content of cultivated strains *T. primitia* ZAS-1, ZAS-2, and *T. azotonutricium* ZAS-9. Collectively these termite *Treponema* possessed GHF enzymes which have specific actives on substrates such as cellulose, chitin, and hemicelluloses but also those with non-specific or other substrate activates (Figure 2A). Although the number of these encoded GHF enzymes varied between these *Treponema*, they were enriched for hemicellulases and GHF enzymes with non-specific activities compared to cellulases and chitinases (Figure 2B). Across the ectosymbiotic communities of *Treponema* associated to either *Trichonympha* or *Dinenympha* hosts, our (meta)transcriptome analysis indicated that they expressed GHF enzymes related to cellulase, hemicellulase, and non-specific hydrolases (Figure 3A).

The gene content related to the Wood-Ljungdahl pathway and nitrogenase were compared across these termite *Treponema* (Figure 2C). *Treponema* bin DS12 encoded all genes belonging to the Wood-Ljungdahl pathway as did the cultivated acetogens *T. primitia* and *T. azotonutricium* (Figure 2C). However, the two genomes in *Treponema* bin DS12 did not contain the *nifHDK* genes encoding the subunits of nitrogenase, which were present in other termite *Treponema* species (Figure 2C). The expression of these genes was investigated across the *Treponema* communities associated with *T. agilis* and *Dinenympha* species hosts (Figure 3B). This analysis indicated that genes involved in reductive acetogenesis were expressed in every sample, however only two samples (TA26 and DS12) had reads mapping to every gene and most samples were missing expression of at least one gene in this pathway (Figure 3B). All three

subunits of nitrogenase (*nifD*, *nifK*, and *nifH*) were expressed in sample DS12 but not in other *Dinenympha* samples (Figure 3B). In *T. agilis* samples only *nifD* and *nifH* were observed to be expressed (Figure 3B). Overall, there was no difference in the expression levels of genes related to the Wood-Ljungdahl pathway or nitrogenase between the *Treponema* communities of *T. agilis* and *Dinenympha* hosts (Mann Whitney Test, $p > 0.05$).

Protein mediated attachment of ectosymbionts and putative binding proteins

Similar to what had been observed by others (Radek and Nitsch, 2007), our TEM analysis revealed that the interface of binding sites between ectosymbiotic *Treponema* and their protist hosts contained electron dense materials (Figure 4A and 4B). Of the chemical treatments used to disrupt these attachments, only treatments with proteinase K detached ectosymbionts from both live and fixed protist cells (Table 1 and Figures 4C – 4F). Our SEM analysis of these treated samples revealed that the detachment was disrupted at the binding interface and resulted in: i) ectosymbiont which were visibly in the process of detaching (Figure 4D), ii) empty binding sites (Figure 4E), iii) and ectosymbionts with blunt ends (Figure 4F) consistent with them being previously adhered to a protist host. These results supported previous studies which concluded that ectosymbionts are attached to their protist host by a protein mediated mechanism (Radek and Nitsch, 2007). Genes that encoded putative cell-cell binding proteins were identified in the *Treponema* bin DS12 genomic bin and compared with those found in *T. primitia* ZAS-1 and ZAS-2 and '*Ca*. Symbiothrix' (Figure 4G). All genomes possessed cell surface proteins that contained leucine rich repeats (LRRs), which have been shown to be involved in binding or adhesion of bacterial symbionts to host cells (Inagaki et al., 2006; Marino et al., 1999). In addition, all of these genomes possessed genes that encoded fibronectin type III domain containing proteins (Figure 4G). Interestingly, only ectosymbiotic *Treponema* bin DS12 and '*Ca*.

Symbiothrix' contained genes that were initially annotated as internalin family proteins (Figure 4G), however after further inspection showed that these do not resemble other internalin family proteins and contain different domains from one another (Supplemental Figure 2).

Discussion

The use single protist cells as template for both assembly and read-based omics analyses of the ectosymbiotic *Treponema* communities gave insights into some of the possible functions these bacteria may provide in the hindguts of *R. flavipes*. The genomes in *Treponema* bin DS12, revealed that they, like their cultivated relatives, likely contributes to reductive acetogenesis in the hindgut of their termite host. In addition, *Treponema* bin DS12 possessed many genes that encoded GHF enzymes suggesting that these *Treponema* may also participate in polysaccharide hydrolysis. This was supported by a metatranscriptome analysis which showed that the communities of ectosymbiotic *Treponema* did express GHF enzymes while attached to their protist hosts. The metatranscriptome analysis also supported that these communities expressed genes related to reductive acetogenesis and nitrogenase activity.

The observations that ectosymbionts which represent two different bacterial taxa (*Treponema* and '*Ca*. Symbiothrix') both possessed genes that encoded GHF enzymes suggests that polysaccharide hydrolysis may be a conserved function of these ectosymbionts. The polymerization state of these polysaccharides is not known. It is also not known if they are degraded for the sole benefit of the bacteria or for the benefit of the host protist as well. In other wood-feeding termites which lack protists, *Treponema* species which are associate with wood fibers have recently been shown to be the major source of hemicellulase activity (Tokuda et al.,

2018). Together, these findings suggest that bacterial members of the hindgut community, may contribute to some of the hydrolysis functions that have traditionally been attributed to the protists in the hindguts of wood-feeding termites (Sethi et al., 2013; Yamin, 1981; Yamin and Trager, 1979). It may be that polysaccharides found in wood are hydrolyzed by various members of the microbial community of wood-feeding termites. The ectosymbionts may either hydrolyze different polysaccharides compared to their hosts or their enzymes may act as a pre-treatment prior to being hydrolyzed by the protists (Yuki et al., 2015).

The previous observations that some members of the termite *Treponema* communities are acetogens had suggested that this function could be conserved across different *Treponema* species which inhabit wood-feeding termites (Graber and Breznak, 2004; Pester and Brune, 2006). Our analyses of the ectosymbiotic communities supports this hypothesis by demonstrating that the ectosymbionts of protists possessed and transcribed genes belonging to the Wood-Ljungdahl pathway for reductive acetogenesis. However, gene content data is currently only available for *Treponema* which belongs to the termite *Treponema* Cluster I, which includes *Treponema primitia* and the two genomes in *Treponema* bin DS12. Our metatranscriptome analysis could not resolve whether members of Cluster II are also acetogens.

Since nitrogen is limiting in the termite's diet of lignocellulose, bacterial nitrogen fixation is essential to maintaining termites and their microbial communities. Previous studies have concluded that some termite-associated *Treponema* are capable of fixing nitrogen and suggested that these members may play a critical role in the hindguts of termites (Lilburn et al., 2001). We did detect transcription of the gene encoding nitrogenase subunits across the ectosymbiotic communities of protist hosts. It has been hypothesized that the protist hosts may recover nitrogenous compounds by digesting a subpopulation of the bacterial community over

time (Brune, 2014; Yuki et al., 2015). The termite then recovers nitrogen by digesting microbial biomass which is shared among nest mates though proctodeal trophallaxis (Brune, 2014). Thus, the ability of termites to live exclusively on nutrient-poor lignocellulose by wood-feeding likely relies heavily on nitrogenase activity of their hindgut microbiota (Brune, 2014; Ohkuma et al., 1996; Tai et al., 2016).

Our assays to investigate the mode of attachments of these ectosymbionts to protist hosts supported that the attachment is protein mediated, as seen in other studies (Radek and Nitsch, 2007). Since both *Treponema* bin DS12 and '*Ca*. Symbiothrix' possessed fibronectin type III domain containing protein these are good candidates for proteins involved in their binding or adhesion to protists hosts.

The use these -omic approaches, applied to unicellular hosts and their ectosymbiotic *Treponema* communities has been useful in supporting previous hypotheses regarding these yet to be cultivated symbionts, as well generating new ones. In addition, they bring new insights into the complex processes and intimacies that exist in the hindguts of wood-feeding termites. Approaches which generate long reads at high coverage may help increase the resolution of these studies and enable more complete genome and transcriptome recovery.

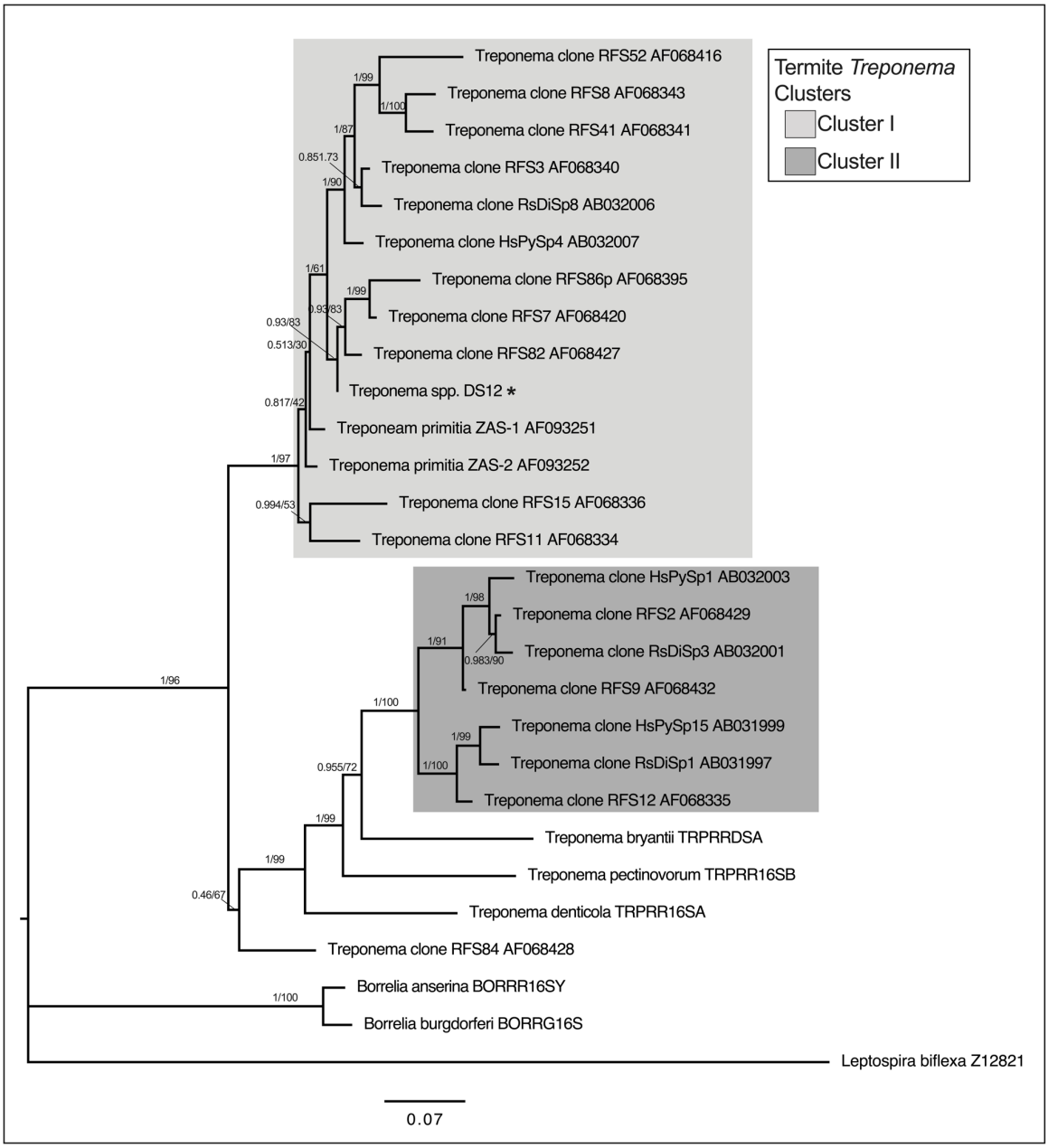


Figure 1 Maximum likelihood (ML) phylogenetic tree of spirochete 16S rRNA genes. The 16S rRNA gene from *Treponema* bin DS12 (denoted by *) was aligned to reference termite associated *Treponema* and other spirochete sequences and a ML tree was generated using the program IQ-Tree with substitution model TN+I+G4. *Treponema* bin DS12 grouped within the Termite *Treponema* Cluster I. Tip labels represent the sequence names and accession numbers. Branch support values represent the Bayesian posterior probability and Bootstrap support values respectively.

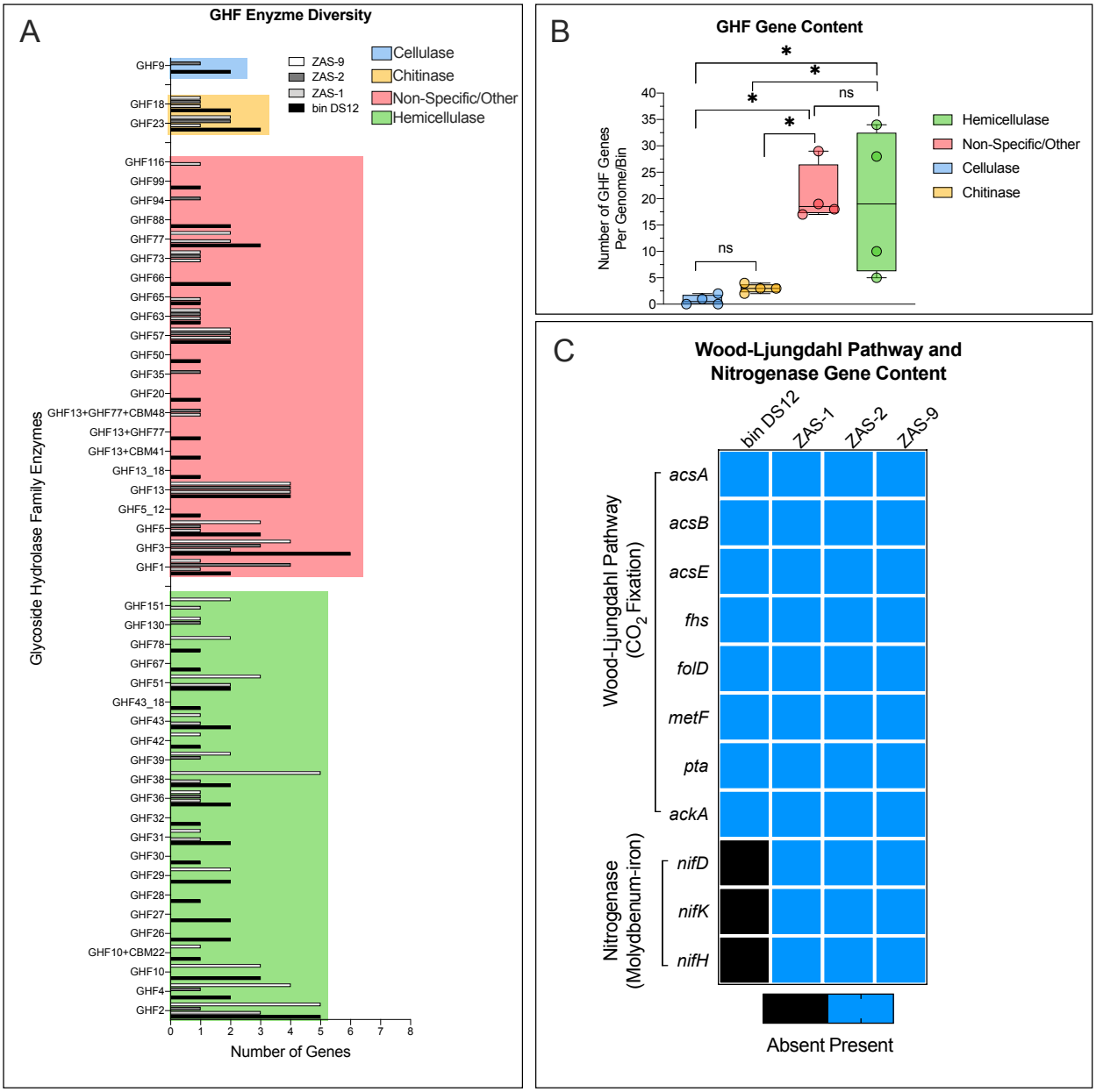


Figure 2 Diversity and gene content of GHF enzymes and genes involved in the Wood-Ljungdahl pathway and nitrogenase in termite-*Treponema*. (A) Diversity of GHF enzymes encoded by *Treponema* bin DS12 compared to cultivated termite *Treponema* strains. Since *Treponema* bin DS12 may contain two genomes, gene counts were divided in half for the comparisons. (B) Comparisons of the number of cellulase, chitinase, hemicellulase, and non-specific/other GHF enzymes encoded by *Treponema* bin DS12 and cultivated termite *Treponema* strains. GHF enzymes were classified and grouped based on their reported activities on the CAZY database. Box plots represent the 25th and 75th percentiles with the line representing the median and whiskers extending to the minimum and maximum values. Statistical significance was tested using Mann-Whitney tests (* $p<0.05$). (C) Presence or absence of genes involved in

the Wood-Ljungdahl-pathway and nitrogenase in *Treponema* bin DS12 and cultivated termite *Treponema* strains.

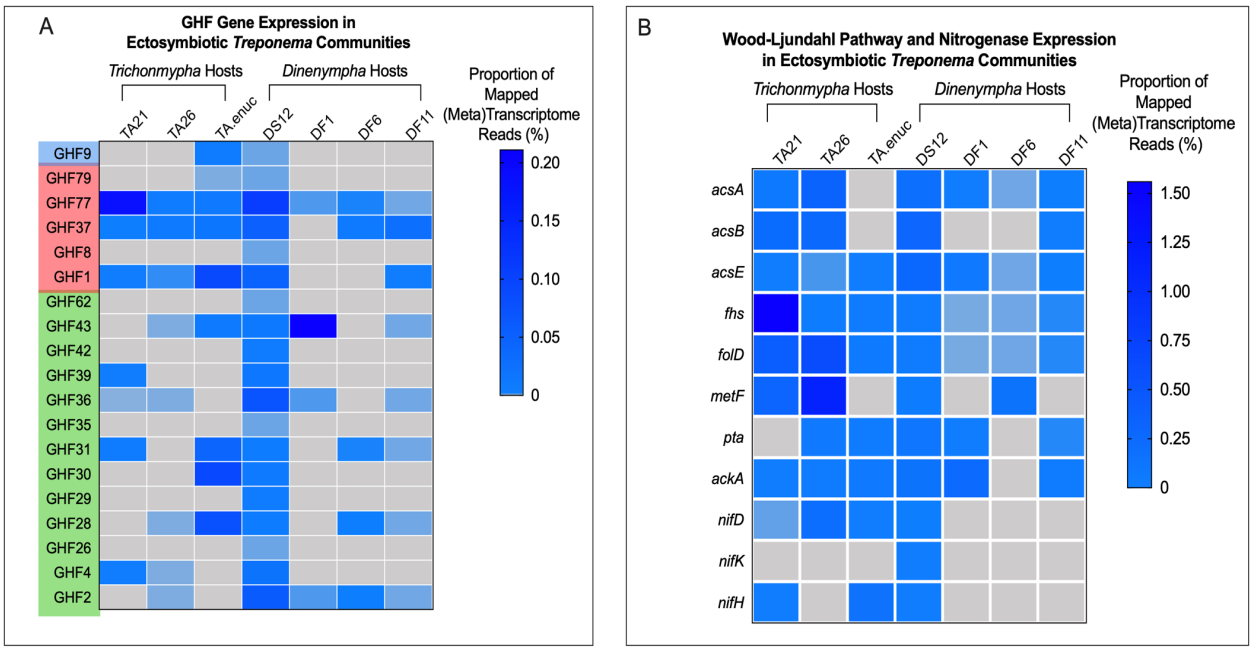
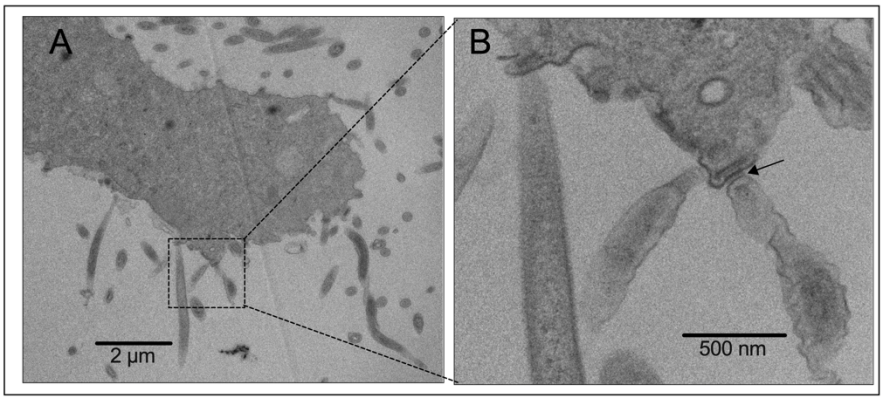
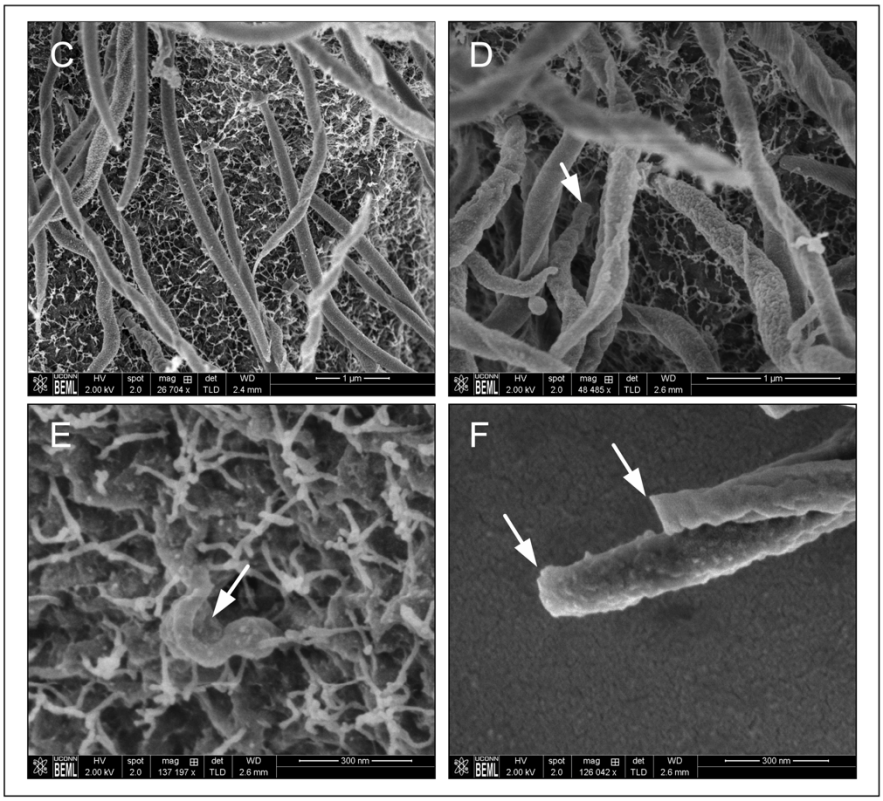


Figure 3 Gene expression of genes involved in carbon and nitrogen fixation and GHF enzymes. (A) Heatmap representing the expression of GHF enzymes by ectosymbiotic *Treponema* communities of individual *Trichonympha* and *Dinenympha* host cells. (B) Heatmap representing the expression of genes involved in the Wood-Ljungdahl pathway and Nitrogenase by ectosymbiotic *Treponema* communities of individual *Trichonympha* and *Dinenympha* host cells.


A
B
2 µm
500 nm


C
D
E
F

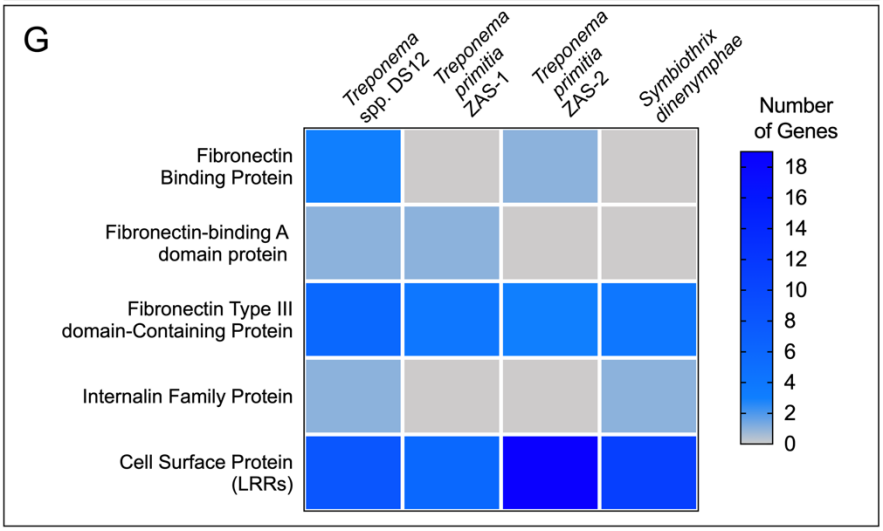

G
Treponema spp. DS12
Treponema primitia ZAS-1
Treponema primitia ZAS-2
Symbiothrix dinenymphae
Fibronectin Binding Protein
Fibronectin-binding A domain protein
Fibronectin Type III domain-Containing Protein
Internalin Family Protein
Cell Surface Protein (LRRs)
Number of Genes
18
16
14
12
10
8
6
4
2
0

Figure 4 Ectosymbiont attachment sites and proteins putatively involved in host cell attachment. (A) TEM micrograph of an ultrathin cross section of a protist host and their attached ectosymbionts. (B) Close up of a binding site showing electron dense materials (arrow) which connects the outer member of the ectosymbiotic bacterium to the plasma membrane of the protist host. (C) SEM micrograph of ectosymbiotic bacteria attached to *a Dinenympha* host cell. (D) SEM micrograph of ectosymbiotic bacteria in the process of detaching from a *Dinenympha* host cell following treatment with proteinase K. (E) SEM micrograph of an empty binding site on a *Dinenympha* host cell following treatment with proteinase K. (F) SEM micrograph of ectosymbionts which have detached from a *Dinenympha* host cell following treatment with proteinase K. (G) Heatmap of the number of putative proteins involved in host binding across different termite *Treponema* and '*Ca*. Symbiothrix'. Proteins that were initially annotated as internalin family proteins were further investigated (Supplemental Figure 2).

Table 1 Chemical perturbations to detach ectosymbionts from protist hosts

Treatment	Concentration	Interaction	Oxymonadida*	Parabasalia*
NaCl	3M		-	-
CaCl2	1M	Ionic	-	-
EDTA	10mM		-	-
Triton-X	0.10%	Hydrophobic	-	-
	0.30%		-	-
Proteinase K	100µg/ml	Protein	+	+
	200µg/ml		+	+

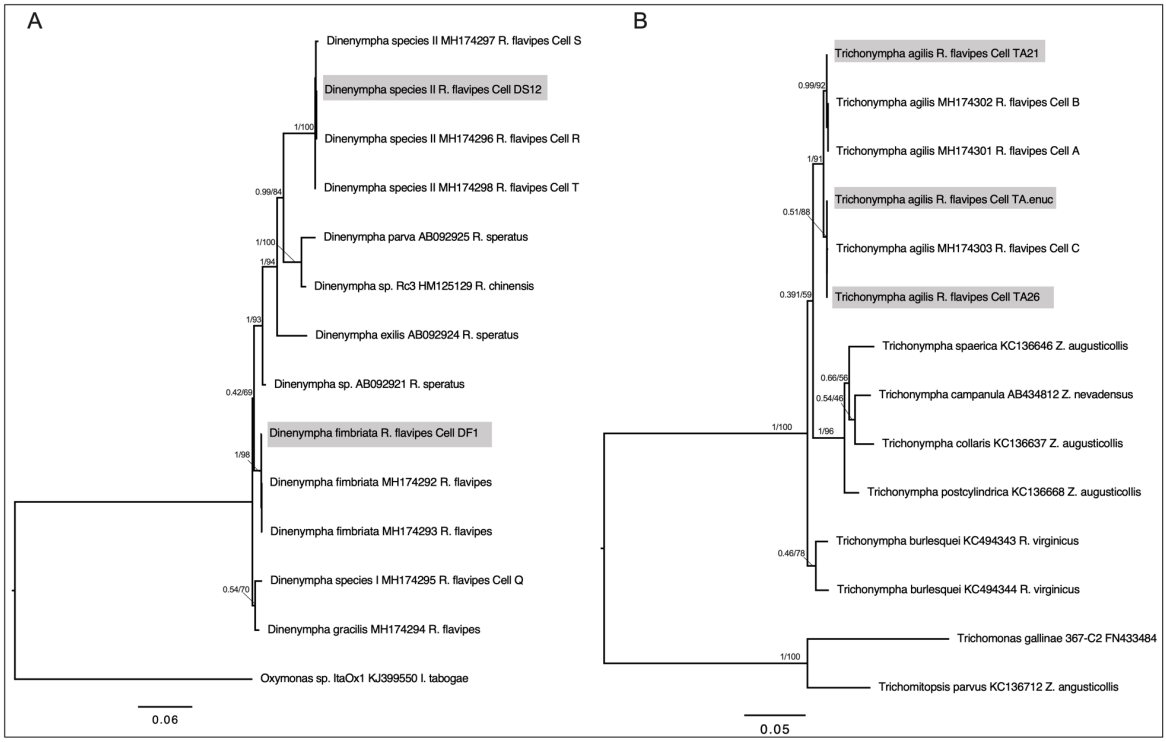
* Represents both live and fixed protist cells

+ Denotes treatments which detached ectosymbionts

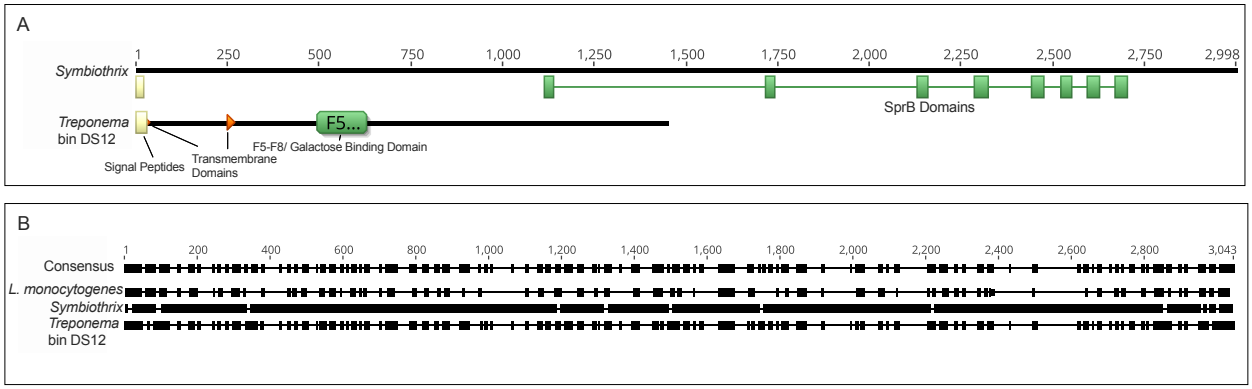
Table 2 Features of termite *Treponema* species Genomes/Genomic Bins

Treponema Genome	Size (Mb)	G+C %	Accession Number
Treponema primitia ZAS-1*	3.5	51	AEEA00000000
Treponema primitia ZAS-2*	3.8	50.9	CP001843
Treponema azotonutricium ZAS-9*	3.9	50	CP001841
Treponema bin DS12	8.0	55.9	-

*Values for genome size, G+C %, and accession number are from reference (Graber et al., 2004)



Supplementary Figure 1 Maximum likelihood (ML) phylogenetic trees of protist 18S rRNA genes. (A) ML tree of *Dinenympha* species used in this study. (B) ML tree of *Trichonympha agilis* and other Parabasilia protists. Taxa in grey denote 18S rRNA genes retrieved from metagenome assemblies of single protist host cells and their symbionts. Support values represent the Bayesian posterior probability and Bootstrap support values respectively.



Supplemental Figure 2 Domain composition and alignment of putative internalin family proteins. (A) Domain composition of putative internalin family proteins found in *Treponema* bin DS12 and '*Ca*. Symbiothrix dinenymphae'. (B) Alignment putative internalin family proteins to *Listeria monocytogenes* InlJ.

Chapter Five

Conclusions and Future Directions

The complex hindgut community of *R. flavipes* contains a multitude of intimate interactions such as those between protists and their associated bacterial symbionts (Brune, 2014; Ohkuma, 2008). Using culture-independent, high-throughput omic assays we focused on these communities to investigate certain principles such as their community structures, transmission trends, and gene content/expression.

The use of high-throughput amplicon sequencing revealed that the bacterial communities of *Dinenympha*, *Pyrsonympha*, and *Trichonympha* hosts were distinct from one another and relatively conserved across termite colonies. In addition, the observation that multiple Spirochete (*Treponema*) ASVs were found associated with all four of the *Dinenympha* species suggested that some ectosymbionts may have associated with multiple hosts species and can be acquired by horizontal transmission. This was supported by an analysis using a novel fluorescent assay to track members of the hindgut community overtime and allowed for the observation that protist hosts acquired newly attached ectosymbiotic bacteria during those assays. This transmission was shown to require active processes and showed regional specificity of some hosts. In addition, some protist-associated *Treponema* ASVs were found in the free-living fraction of bacteria in the termite hindgut. These data collectively suggested that members of the communities of ectosymbiotic bacteria may be horizontally acquired.

The use of metagenomic and metatranscriptomic sequencing and genomic binning was useful in revealing differences in gene content between different endosymbiotic *Endomicrobium* species. This led to the discovery of that these endosymbionts used different wood-derived carbon sources and had other differences related to aerotolerance. These differences also prompted us to investigate whether the genes related to the differences in carbon metabolism could have been acquired by horizontal gene transfer (HGT). These analyses showed each of these carbon usage pathways had been acquired by HGT and from donor taxa that are present in the hindgut community of the termite. Further analysis revealed that these endosymbiotic lineages of *Endomicrobium* possessed the genes that encode proteins that are involved in a known competence pathway in other organisms. These genes were shown to be well conserved in the

Elusimicrobia phylum, and likely to be functional based on a dN/dS analysis. The genes were expressed by the endosymbiotic *Endomicrobium* species. These data have led to the hypothesis that these endosymbionts may be able to become competent and that this may be one mechanism by which they acquire novel genes.

These methods also allowed us to investigate the communities of ectosymbiotic *Treponema* across *Trichonympha* and *Dinenympha* hosts. With these data we tested whether ectosymbionts both possessed and expressed genes related to certain processes that are considered to be essential in termite hindgut. These included polysaccharide hydrolysis, nitrogen fixation, and reductive acetogenesis. Using a *Treponema* genomic bin recovered from a *Dinenympha* host cell as well as a metatranscriptome analysis we showed that members of these communities do possess and transcribe genes related to those functions. In addition, we identified genes that encoded proteins involved in ectosymbiont attachment to protist hosts. These included fibronectin type III domain-containing proteins and cell surface proteins that contained leucine rich repeats (LRRs).

Future directions of this include testing of these precise hypotheses generated from this omics data as well as continuing the use of omics with long read technologies to further obtain genomes and investigate gene expression of these protist-associated communities, For example, the candidate *Treponema* genes that may be involved in cell adhesion could be expressed and the proteins could be used to generate antibodies for immunofluorescence assays. These could then be used to test whether those proteins may be involved in binding. Similarly, the *Endomicrobium* competence genes could be expressed in other organisms such that their functions can be verified. One organism can be used for this would be *Vibrio fischeri* which possesses a similar competence system can be assayed in vitro (Pollack-Berti et al., 2010).

References

Abby, S. S., Néron, B., Ménager, H., Touchon, M., and Rocha, E. P. C. (2014). MacSyFinder: A program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One* 9. doi:10.1371/journal.pone.0110726.

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., et al. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* 36, 566–569. doi:10.1038/nbt.4163.

Aziz, R. K., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics* 9, 1–15. doi:10.1186/1471-2164-9-75.

Benjamino, J., and Graf, J. (2016). Characterization of the Core and Caste-Specific Microbiota in the Termite, Reticulitermes flavipes. *Front. Microbiol.* 7, 171. doi:10.3389/fmicb.2016.00171.

Bhattacharya, D., Yoon, H. S., Price, D. C., Stepanauskas, R., Rajah, V. D., Sieracki, M. E., et al. (2011). Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists. *Science (80-.).* 332, 714–717. doi:10.1126/science.1203163.

Boucias, D. G., Cai, Y., Sun, Y., Lietze, V. U., Sen, R., Raychoudhury, R., et al. (2013). The hindgut lumen prokaryotic microbiota of the termite Reticulitermes flavipes and its responses to dietary lignocellulose composition. *Mol. Ecol.* 22, 1836–1853. doi:10.1111/mec.12230.

Boulos, L., Prévost, M., Barbeau, B., Coallier, J., and Desjardins, R. (1999). LIVE/DEAD(®) BacLight(TM): Application of a new rapid staining method for direct enumeration of viable and total bacteria in drinking water. *J. Microbiol. Methods* 37, 77–86. doi:10.1016/S0167-7012(99)00048-2.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., et al. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* 5. doi:10.1038/srep08365.

Breznak, J. a., and Pankratz, H. S. (1977). In situ morphology of the gut microbiota of wood-eating termites [Reticulitermes flavipes (Kollar) and Coptotermes formosanus Shiraki]. *Appl. Environ. Microbiol.* 33, 406–426.

Bright, M., and Bulgheresi, S. (2010). A complex journey: transmission of microbial symbionts. *Nat. Rev. Microbiol.* 8, 218–230. doi:10.1038/nrmicro2262.

Brune, A. (2014). Symbiotic digestion of lignocellulose in termite guts. *Nat. Rev. Microbiol.* 12, 168–80. doi:10.1038/nrmicro3182.

Brune, A., Emerson, D., and Breznak, J. A. (1995). The termite gut microflora as an oxygen sink: Microelectrode determination of oxygen and pH gradients in guts of lower and higher termites. *Appl. Environ. Microbiol.* 61, 2681–2687.

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using

DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176.

Bushnell, B. (2014). BBMap: A Fast, Accurate, Splice-Aware Aligner. doi:10.1186/1471-2105-13-238.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi:10.1038/ismej.2017.119.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi:10.1038/nmeth.3869.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi:10.1038/ismej.2012.8.

Caruso, V., Song, X., Asquith, M., and Karstens, L. (2019). Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. *mSystems* 4, 1–19. doi:10.1128/msystems.00163-18.

Chouvenc, T., and Foley, J. R. (2018). Coptotermes gestroi (Wasmann) (Blattodea [Isoptera]: Rhinotermitidae), a Threat to the Southeastern Florida Urban Tree Canopy. *Florida Entomol.* 101, 79–90. doi:10.1653/024.101.0115.

Coil, D., Jospin, G., and Darling, A. E. (2015). A5-miseq: An updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics* 31, 587–589. doi:10.1093/bioinformatics/btu661.

Corno, G., and Jürgens, K. (2006). Direct and indirect effects of protist predation on population size structure of a bacterial strain with high phenotypic plasticity. *Appl. Environ. Microbiol.* 72, 78–86. doi:10.1128/AEM.72.1.78.

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6, 226. doi:10.1186/s40168-018-0605-2.

de Goffau, M. C., Lager, S., Salter, S. J., Wagner, J., Kronbichler, A., Charnock-Jones, D. S., et al. (2018). Recognizing the reagent microbiome. *Nat. Microbiol.* 3, 851–853. doi:10.1038/s41564-018-0202-y.

Douglas, A. E. (2018). Omics and the metabolic function of insect–microbial symbioses. *Curr. Opin. Insect Sci.* 29, 1–6. doi:10.1016/j.cois.2018.05.012.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi:10.1093/nar/gkh340.

Ellison, C. K., Dalia, T. N., Vidal Ceballos, A., Wang, J. C. Y., Biais, N., Brun, Y. V., et al.

(2018). Retraction of DNA-bound type IV competence pili initiates DNA uptake during natural transformation in Vibrio cholerae. *Nat. Microbiol.* 3, 773–780. doi:10.1038/s41564-018-0174-y.

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi:10.1111/2041-210X.12114.

Evangelista, D. A., Wipfler, B., Simon, S., Béthoux, O., Donath, A., Fujita, M., et al. (2019). An integrative phylogenomic approach illuminates the evolutionary history of cockroaches and termites (Blattodea) Beijing Advanced Innovation Center for Food Nutrition and Human Health, and. doi:10.1098/rspb.2018.2076.

Evans, T. A., Dawes, T. Z., Ward, P. R., and Lo, N. (2011). Ants and termites increase crop yield in a dry climate. *Nat. Commun.* 2, 262–267. doi:10.1038/ncomms1257.

Graber, J. R., and Breznak, J. a (2004). Physiology and Nutrition of Treponema primitia , an H 2 / CO 2 -Acetogenic Spirochete from Termite Hindguts Physiology and Nutrition of Treponema primitia , an H 2 / CO 2 -Acetogenic Spirochete from Termite Hindguts. *Appl. Environ. Microbiol.* 70, 1307–1314. doi:10.1128/AEM.70.3.1307.

Graber, J. R., Leadbetter, J. a. R., and Breznak, J. a. (2004). Description of *Treponema azotonutricium* sp. nov. and *Treponema primitia* sp. nov., the first spirochetes isolated from termite guts. *Appl. Environ. Microbiol.* 70, 1315–1320. doi:10.1128/AEM.70.3.1315.

Hamann, E., Tegetmeyer, H. E., Riedel, Di., Littmann, S., Ahmerkamp, S., Chen, J., et al. (2017). Syntrophic linkage between predatory Carpediemonas and specific prokaryotic populations. *ISME J.* 11, 1205–1217. doi:10.1038/ismej.2016.197.

Hongoh, Y., Sato, T., Noda, S., Ui, S., Kudo, T., and Ohkuma, M. (2007). Candidatus Symbiothrix dinenymphae: Bristle-like Bacteroidales ectosymbionts of termite gut protists. *Environ. Microbiol.* 9, 2631–2635. doi:10.1111/j.1462-2920.2007.01365.x.

Hongoh, Y., Sharma, V. K., Prakash, T., Noda, S., Taylor, T. D., Kudo, T., et al. (2008a). Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *PNAS* 105, 5555–5560.

Hongoh, Y., Sharma, V. K., Prakash, T., Noda, S., Toh, H., Taylor, T. D., et al. (2008b). Genome of an endosymbiont coupling N2 fixation to cellulolysis within protist cells in termite gut. *Science* 322, 1108–9. doi:10.1126/science.1165578.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* 12, 1–12. doi:10.1371/journal.pcbi.1004957.

Iida, T., Ohkuma, M., Ohtoko, K., and Kudo, T. (2000). Symbiotic spirochetes in the termite hindgut: Phylogenetic identification of ectosymbiotic spirochetes of oxymonad protists. *FEMS Microbiol. Ecol.* 34, 17–26. doi:10.1016/S0168-6496(00)00070-2.

Ikeda-Ohtsubo, W., and Brune, A. (2009). Cospeciation of termite gut flagellates and their bacterial endosymbionts: Trichonympha species and "Candidatus Endomicrobium trichonymphae." *Mol. Ecol.* 18, 332–342. doi:10.1111/j.1365-294X.2008.04029.x.

Ikeda-Ohtsubo, W., Desai, M., Stingl, U., and Brune, A. (2007). Phylogenetic diversity of "Endomicrobia" and their specific affiliation with termite gut flagellates. *Microbiology* 153, 3458–3465. doi:10.1099/mic.0.2007/009217-0.

Ikeda-Ohtsubo, W., Faivre, N., and Brune, A. (2010). Putatively free-living 'Endomicrobia'-ancestors of the intracellular symbionts of termite gut flagellates? *Environ. Microbiol. Rep.* 2, 554–9. doi:10.1111/j.1758-2229.2009.00124.x.

Inagaki, S., Onishi, S., Kuramitsu, H. K., and Sharma, A. (2006). Porphyromonas gingivalis vesicles enhance attachment, and the leucine-rich repeat BspA protein is required for invasion of epithelial cells by "Tannerella forsythia." *Infect. Immun.* 74, 5023–5028. doi:10.1128/IAI.00062-06.

Ishak, H. D., Miller, J. L., Sen, R., Dowd, S. E., Meyer, E., and Mueller, U. G. (2011). Microbiomes of ant castes implicate new microbial roles in the fungus-growing ant Trachymyrmex septentrionalis Heather. *Sci. Rep.* 1, 20–22. doi:10.1038/srep00204.

Izawa, K., Kuwahara, H., Kihara, K., Yuki, M., Lo, N., Itoh, T., et al. (2016). Comparison of intracellular "ca. endomicrobium trichonymphae" genomovars illuminates the requirement and decay of defense systems against foreign DNA. *Genome Biol. Evol.* 8, 3099–3107. doi:10.1093/gbe/evw227.

Izawa, K., Kuwahara, H., Sugaya, K., Lo, N., Ohkuma, M., and Hongoh, Y. (2017). Discovery of ectosymbiotic Endomicrobium lineages associated with protists in the gut of stolotermitid termites. *Environ. Microbiol. Rep.* 9, 411–418. doi:10.1111/1758-2229.12549.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 1–8. doi:10.1038/s41467-018-07641-9.

James, E. R., Tai, V., Scheffrahn, R. H., and Keeling, P. J. (2013). Trichonympha burlesquei n. sp. From Reticulitermes virginicus and evidence against a cosmopolitan distribution of Trichonympha agilis in many termite hosts. *Int. J. Syst. Evol. Microbiol.* 63, 3873–3876. doi:10.1099/ijs.0.054874-0.

Jasso-Selles, D. E., De Martini, F., Freeman, K. D., Garcia, M. D., Merrell, T. L., Scheffrahn, R. H., et al. (2017). The parabasalid symbiont community of Heterotermes aureus: Molecular and morphological characterization of four new species and reestablishment of the genus Cononympha. *Eur. J. Protistol.* 61, 48–63. doi:10.1016/j.ejop.2017.09.001.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30.

Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S. C., Petrželková, R., Eme, L., et al. (2016). A eukaryote without a mitochondrial organelle. *Curr. Biol.* 26, 1274–1284. doi:10.1016/j.cub.2016.03.053.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi:10.1093/bioinformatics/bts199.

Kerwin, A. H., and Nyholm, S. V (2018). Reproductive System Symbiotic Bacteria Are Conserved Oahu , Hawaii. *msphere* 3, 1–9.

Kodama, Y., and Fujishima, M. (2008). Cycloheximide Induces Synchronous Swelling of Perialgal Vacuoles Enclosing Symbiotic Chlorella vulgaris and Digestion of the Algae in the Ciliate Paramecium bursaria. *Protist* 159, 483–494. doi:10.1016/j.protis.2008.02.005.

Kolisko, M., Boscaro, V., Burki, F., Lynn, D. H., and Keeling, P. J. (2014). Single-cell transcriptomics for microbial eukaryotes. *Curr. Biol.* 24, R1081–R1082. doi:10.1016/j.cub.2014.10.026.

Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H. H., et al. (2015). VizBin - An application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome* 3, 1–7. doi:10.1186/s40168-014-0066-1.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923.

Leadbetter, J. R., Schmidt, T. M., Graber, J. R., and Breznak, J. a (1999). Acetogenesis from H2 plus CO2 by spirochetes from termite guts. *Science* 283, 686–689. doi:10.1126/science.283.5402.686.

Lewis, J. L., and Forschler, B. T. (2004). Protist communities from four castes and three species of Reticulitermes (Isoptera : Rhinotermitidae). *Ann. Entomol. Soc. Am.* 97, 1242–1251. doi:10.1603/0013-8746(2004)097[1242:pcffca]2.0.co;2.

Lewis, J. L., and Forschler, B. T. (2006). A nondichotomous key to protist species identification of *Reticulitermes* (Isoptera: Rhinotermitidae). *Ann. Entomol. Soc. Am.* 99, 1028–1033. doi:10.1603/0013-8746(2006)99.

Lilburn, T. G., Kim, K. S., Ostrom, N. E., Byzek, K. R., Leadbetter, J. R., and Breznak, J. a (2001). Nitrogen fixation by symbiotic and free-living spirochetes. *Science* 292, 2495–2498. doi:10.1126/science.1060281.

Liu, N., Li, H., Chevrette, M. G., Zhang, L., Cao, L., Zhou, H., et al. (2019). Functional metagenomics reveals abundant polysaccharide-degrading gene clusters and cellobiose utilization pathways within gut microbiota of a wood-feeding higher termite. *ISME J.* 13, 104–117. doi:10.1038/s41396-018-0255-1.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, 490–495. doi:10.1093/nar/gkt1178.

López-Madrigal, S., and Gil, R. (2017). Et tu, brute? Not even intracellular mutualistic symbionts escape horizontal gene transfer. *Genes (Basel).* 8, 1–16. doi:10.3390/genes8100247.

Lumppio, H., Shenvi, N. V., Summers, A. O., Voordouw, G., and Kurtz JR., D. M. (2001). Rubrerythrin and Rubredoxin Oxidoreductase in Desulfovibrio vulgaris. *Society* 183, 101–108. doi:10.1128/JB.183.1.101.

Lux, R., Miller, J. N., and Park, N. (2001). Motility and Chemotaxis in Tissue Penetration of Oral Epithelial Cell Layers by Treponema denticola. 69, 6276–6283. doi:10.1128/IAI.69.10.6276.

Maekawa, K., Lo, N., Kitade, O., Miura, T., and Matsumoto, T. (1999). Molecular phylogeny and geographic distribution of wood-feeding cockroaches in East Asian Islands. *Mol. Phylogenet. Evol.* 13, 360–376. doi:10.1006/mpev.1999.0647.

Mangot, J. F., Logares, R., Sánchez, P., Latorre, F., Seeleuthner, Y., Mondy, S., et al. (2017). Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci. Rep.* 7, 1–12. doi:10.1038/srep41498.

Marino, M., Braun, L., Cossart, P., and Ghosh, P. (1999). Structure of the InIB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen L. monocytogenes. *Mol. Cell* 4, 1063–1072. doi:10.1016/S1097-2765(00)80234-8.

McFall-Ngai, M. J. (2014). The Importance of Microbes in Animal Development: Lessons from the Squid-Vibrio Symbiosis. *Annu. Rev. Microbiol.* doi:10.1146/annurev-micro-091313-103654.

McMurdie, P. J., and Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* 8. doi:10.1371/journal.pone.0061217.

Mikaelyan, A., Thompson, C. L., Meuser, K., Zheng, H., Rani, P., Plarre, R., et al. (2017). High-resolution phylogenetic analysis of Endomicrobia reveals multiple acquisitions of endosymbiotic lineages by termite gut flagellates. *Environ. Microbiol. Rep.* 9, 477–483. doi:10.1111/1758-2229.12565.

Mitchell, A., Chang, H. Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. (2015). The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Res.* 43, D213–D221. doi:10.1093/nar/gku1243.

Nauer, P. A., Hutley, L. B., and Arndt, S. K. (2018). Termite mounds mitigate half of termite methane emissions. *Proc. Natl. Acad. Sci.* 115, 13306–13311. doi:10.1073/pnas.1809790115.

Nayfach, S., and Pollard, K. S. (2016a). Toward Accurate and Quantitative Comparative Metagenomics. *Cell* 166, 1103–1116. doi:10.1016/j.cell.2016.08.007.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi:10.1093/molbev/msu300.

Noda, S., Inoue, T., Hongoh, Y., Kawai, M., Nalepa, C. a., Vongkaluang, C., et al. (2006). Identification and characterization of ectosymbionts of distinct lineages in Bacteroidales attached to flagellated protists in the gut of termites and a wood-feeding cockroach. *Environ. Microbiol.* 8, 11–20. doi:10.1111/j.1462-2920.2005.00860.x.

Noda, S., Ohkuma, M., Yamada, A., Hongoh, Y., and Kudo, T. (2003). Phylogenetic position and in situ identification of ectosymbiotic spirochetes on protists in the termite gut. *Appl. Environ. Microbiol.* 69, 625–633. doi:10.1128/AEM.69.1.625-633.2003.

Ohkuma, M. (2008). Symbioses of flagellates and prokaryotes in the gut of lower termites. *Trends Microbiol.* 16, 345–352. doi:10.1016/j.tim.2008.04.004.

Ohkuma, M., Noda, S., Hattori, S., Iida, T., Yuki, M., Starns, D., et al. (2015). Acetogenesis from H_2 plus CO_2 and nitrogen fixation by an endosymbiotic spirochete of a termite-gut cellulolytic protist. *Proc. Natl. Acad. Sci.* 112, 10224–10230. doi:10.1073/pnas.1423979112.

Ohkuma, M., Noda, S., Usami, R., Horikoshi, K., and Kudo, T. (1996). Diversity of Nitrogen Fixation Genes in the Symbiotic Intestinal Microflora of the Termite Reticulitermes speratus. *Appl. Environ. Microbiol.* 62, 2747–2752.

Ohkuma, M., Ohtoko, K., Grunau, C., Moriya, S., and Kudo, T. (1998). Phylogenetic identification of the symbiotic hypermastigote Trichonympha agilis in the hindgut of the termite Reticulitermes speratus based on small-subunit rRNA sequence. *J. Eukaryot. Microbiol.* 45, 439–444. doi:10.1111/j.1550-7408.1998.tb05096.x.

Ohkuma, M., Sato, T., Noda, S., Ui, S., Kudo, T., and Hongoh, Y. (2007). The candidate phylum "Termite Group 1" of bacteria: Phylogenetic diversity, distribution, and endosymbiont members of various gut flagellated protists. *FEMS Microbiol. Ecol.* 60, 467–476. doi:10.1111/j.1574-6941.2007.00311.x.

Oksanen, A. J., Kindt, R., Legendre, P., Hara, B. O., Simpson, G. L., Stevens, M. H. H., et al. (2008). The vegan Package; Community Ecology Package (Version 1.15-1).

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from. *Cold Spring Harb. Lab. Press Method* 1, 1–31. doi:10.1101/gr.186072.114.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 903, 1–10. doi:10.1038/s41564-017-0012-7.

Pedro, M. a De, Grünfelder, C. G., and Gru, C. G. (2004). Restricted Mobility of Cell Surface Proteins in the Polar Regions of Escherichia coli Restricted Mobility of Cell Surface Proteins in the Polar Regions of Escherichia coli. *Society* 186. doi:10.1128/JB.186.9.2594.

Pester, M., and Brune, A. (2006). Expression profiles of fhs (FTHFS) genes support the hypothesis that spirochaetes dominate reductive acetogenesis in the hindgut of lower termites. *Environ. Microbiol.* 8, 1261–1270. doi:10.1111/j.1462-2920.2006.01020.x.

Peterson, B. F., Stewart, H. L., and Scharf, M. E. (2015). Quantification of symbiotic contributions to lower termite lignocellulose digestion using antimicrobial treatments. *Insect Biochem. Mol. Biol.* 59, 80–88. doi:10.1016/j.ibmb.2015.02.009.

Pimentel, Z. T., and Zhang, Y. (2018). Evolution of the natural transformation protein, ComEC, in Bacteria. *Front. Microbiol.* 9, 1–10. doi:10.3389/fmicb.2018.02980.

Pollack-Berti, A., Wollenberg, M. S., and Ruby, E. G. (2010). Natural transformation of Vibrio fischeri requires tfoX and tfoY. *Environ. Microbiol.* 12, 2302–2311. doi:10.1111/j.1462-2920.2010.02250.x.

Pontes, M. H., Babst, M., Lochhead, R., Oakeson, K., Smith, K., and Dale, C. (2008). Quorum sensing primes the oxidative stress response in the insect endosymbiont, Sodalis glossinidius. *PLoS One* 3. doi:10.1371/journal.pone.0003541.

Pramono, A. K., Kuwahara, H., Itoh, T., Toyoda, A., Yamada, A., and Hongoh, Y. (2017). Discovery and Complete Genome Sequence of a Bacteriophage from an Obligate Intracellular Symbiont of a Cellulolytic Protist in the Termite Gut. *Microbes Environ. Environ.* 32, 112–117. doi:10.1264/jsme2.me16175.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, 61–65. doi:10.1093/nar/gkl842.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* 41, 590–596. doi:10.1093/nar/gks1219.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 833–844. doi:10.1038/nbt.3935.

Radek, R., and Nitsch, G. (2007). Ectobiotic spirochetes of flagellates from the termite Mastotermes darwiniensis: Attachment and cyst formation. *Eur. J. Protistol.* 43, 281–294. doi:10.1016/j.ejop.2007.06.004.

Radek, R., and Tischendorf, G. (1999). Bacterial adhesion to different termite flagellates: Ultrastructural and functional evidence for distinct molecular attachment modes. *Protoplasma* 207, 43–53. doi:10.1007/BF01294712.

Richter, M., Rosselló-Móra, R., Oliver Glöckner, F., and Peplies, J. (2016). JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 32, 929–931. Available at: http://dx.doi.org/10.1093/bioinformatics/btv681.

Sangwan, N., Xia, F., and Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4, 8. doi:10.1186/s40168-016-0154-5.

Sato, T., Hongoh, Y., Noda, S., Hattori, S., Ui, S., and Ohkuma, M. (2009). Candidatus Desulfovibrio trichonymphae, a novel intracellular symbiont of the flagellate Trichonympha agilis in termite gut. *Environ. Microbiol.* 11, 1007–15. doi:10.1111/j.1462-2920.2008.01827.x.

Sato, T., Kuwahara, H., Fujita, K., Noda, S., Kihara, K., Yamada, A., et al. (2014). Intranuclear verrucomicrobial symbionts and evidence of lateral gene transfer to the host protist in the termite gut. *ISME J.* 8, 1008–19. doi:10.1038/ismej.2013.222.

Scharf, M. E. (2014). Termites as Targets and Models for Biotechnology. *Annu. Rev. Entomol.* 60, 77–102. doi:10.1146/annurev-ento-010814-020902.

Scharf, M. E. (2015). Omic research in termites: An overview and a roadmap. *Front. Genet.* 6. doi:10.3389/fgene.2015.00076.

Seitz, P., and Blokesch, M. (2013). DNA-uptake machinery of naturally competent Vibrio

cholerae. *Proc. Natl. Acad. Sci.* 110, 17987–17992. doi:10.1073/pnas.1315647110.

Sethi, A., Kovaleva, E. S., Slack, J. M., Brown, S., Buchman, G. W., and Scharf, M. E. (2013). A GHF7 cellulase from the protist symbiont community of reticulitermes flavipes enables more efficient lignocellulose processing by host enzymes. *Arch. Insect Biochem. Physiol.* 84, 175–193. doi:10.1002/arch.21135.

Stephens, M. E., and Gage, D. J. (2018). Oligotyping Bacterial Symbionts of Hindgut Protists Reveals Complex Population Structures and Transmission Trends in the Termite Reticulitermes flavipes. *bioRxiv*, 299255. doi:10.1101/299255.

Stingl, U., Radek, R., Yang, H., and Brune, A. (2005). "Endomicrobia": Cytoplasmic symbionts of termite gut protozoa form a separate phylum of prokaryotes. *Appl. Environ. Microbiol.* 71, 1473–1479. doi:10.1128/AEM.71.3.1473-1479.2005.

Strassert, J. F. H., Mikaelyan, A., Woyke, T., and Brune, A. (2016). Genome analysis of ' Candidatus Ancillula trichonymphae'', first representative of a deep-branching clade of Bifidobacteriales , strengthens evidence for convergent evolution in flagellate endosymbionts.' *Environ. Microbiol. Rep.* 8, 865–873. doi:10.1111/1758-2229.12451.

Su, N. Y., and Scheffrahn, R. H. (1998). A review of subterranean termite control practices and prospects for integrated pest management programmes. *Integr. Pest Manag. Rev.* 3, 1–13. doi:10.1023/A:1009684821954.

Tai, V., Carpenter, K. J., Weber, P. K., Nalepa, C. A., Perlman, S. J., and Keeling, P. J. (2016). Genome evolution and nitrogen-fixation in bacterial ectosymbionts of a protist inhabiting wood-feeding cockroaches. *Appl. Environ. Microbiol.*, in press. doi:10.1128/AEM.00611-16.

Tang, Y. P., Dallas, M. M., and Malamy, M. H. (1999). Characterization of the BatI (Bacteroides aerotolerance) operon in Bacteroides fragilis: Isolation of a B. fragilis mutant with reduced aerotolerance and impaired growth in in vivo model systems. *Mol. Microbiol.* 32, 139–149. doi:10.1046/j.1365-2958.1999.01337.x.

Tikhonenkov, D. V., Janouškovec, J., Keeling, P. J., and Mylnikov, A. P. (2016). The Morphology, Ultrastructure and SSU rRNA Gene Sequence of a New Freshwater Flagellate, Neobodo borokensis n. sp. (Kinetoplastea, Excavata). *J. Eukaryot. Microbiol.* 63, 220–232. doi:10.1111/jeu.12271.

Tokuda, G., Mikaelyan, A., Fukui, C., Matsuura, Y., Watanabe, H., Fujishima, M., et al. (2018). Fiber-associated spirochetes are major agents of hemicellulose degradation in the hindgut of wood-feeding higher termites. *Proc. Natl. Acad. Sci.* 115, E11996–E12004. doi:10.1073/pnas.1810550115.

Trager, W. (1934). The Cultivation of a Cellulose-Digesting Flagellate, Trichomonas Termopsidis, and of Certain other Termite Protozoa. *Biol. Bull.* 66, 182–190.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40, 1–12. doi:10.1093/nar/gks596.

Utami, Y. D., Kuwahara, H., Igai, K., Murakami, T., Sugaya, K., Morikawa, T., et al. (2019).

Genome analyses of uncultured TG2/ZB3 bacteria in 'Margulisbacteria' specifically attached to ectosymbiotic spirochetes of protists in the termite gut. *ISME J.* 13, 455–467. doi:10.1038/s41396-018-0297-4.

Vacek, V., Novák, L. V. F., Treitli, S. C., Táborský, P., Cepicka, I., Kolísko, M., et al. (2018). Fe-S Cluster Assembly in Oxymonads and Related Protists. *Mol. Biol. Evol.* 35, 2712–2718. doi:10.1093/molbev/msy168.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9. doi:10.1371/journal.pone.0112963.

Wickham, H. (2011). Ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* 3, 180–185. doi:10.1002/wics.147.

Xu, B., and Yang, Z. (2013). PamlX: A graphical user interface for PAML. *Mol. Biol. Evol.* 30, 2723–2724. doi:10.1093/molbev/mst179.

Yamin, M. a., and Trager, W. (1979). Cellulolytic Activity of an Axenically-cultivated Termite Flagellate, Trichomitopsis termopsidis. *J. Gen. Microbiol.* 113, 417–420. doi:10.1099/00221287-113-2-417.

Yamin, M. a (1981). Cellulose metabolism by the flagellate trichonympha from a termite is independent of endosymbiotic bacteria. *Science* 211, 58–59. doi:10.1126/science.211.4477.58.

Yang, H., Schmitt-Wagner, D., Stingl, U., and Brune, A. (2005). Niche heterogeneity determines bacterial community structure in the termite gut (Reticulitermes santonensis). *Environ. Microbiol.* 7, 916–932. doi:10.1111/j.1462-2920.2005.00760.x.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* 13, 555–556. doi:10.1093/bioinformatics/13.5.555.

Ye, W., Lee, C. Y., Scheffrahn, R. H., Aleong, J. M., Su, N. Y., Bennett, G. W., et al. (2004). Phylogenetic relationships of nearctic Reticulitermes species (Isoptera: Rhinotermitidae) with particular reference to Reticulitermes arenincola Goellner. *Mol. Phylogenet. Evol.* 30, 815–822. doi:10.1016/S1055-7903(03)00230-6.

Yuki, M., Kuwahara, H., Shintani, M., Izawa, K., Sato, T., Starns, D., et al. (2015). Dominant ectosymbiotic bacteria of cellulolytic protists in the termite gut also have the potential to digest lignocellulose. *Environ. Microbiol.* 17, 4942–4953. doi:10.1111/1462-2920.12945.

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., et al. (2018). DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 46, W95–W101. doi:10.1093/nar/gky418.

Zheng, H., Dietrich, C., and Brune, A. (2017). Genome analysis of Endomicrobium proavitum suggests loss and gain of relevant functions during the evolution of intracellular symbionts. *Appl. Environ. Microbiol.* 83, 1–14. doi:10.1128/AEM.00656-17.

Zheng, H., Dietrich, C., Hongoh, Y., and Brune, A. (2016a). Restriction-modification systems as mobile genetic elements in the evolution of an intracellular symbiont. *Mol. Biol. Evol.* 33,

721–725. doi:10.1093/molbev/msv264.

Zheng, H., Dietrich, C., L. Thompson, C., Meuser, K., and Brune, A. (2015). Population Structure of Endomicrobia in Single Host Cells of Termite Gut Flagellates (*Trichonympha* spp.). *Microbes Environ.* 30, 92–98. doi:10.1264/jsme2.ME14169.

Zheng, H., Dietrich, C., Radek, R., and Brune, A. (2016b). Endomicrobium proavitum , the first isolate of Endomicrobia class . nov . (phylum Elusimicrobia) – an ultramicrobacterium with an unusual cell cycle that fixes nitrogen with a Group IV nitrogenase. *Environ. Microbiol.* 18, 191–204. doi:10.1111/1462-2920.12960.